Année 2000

MÉMOIRE

Présenté devant

l'Université Claude Bernard - Lyon 1

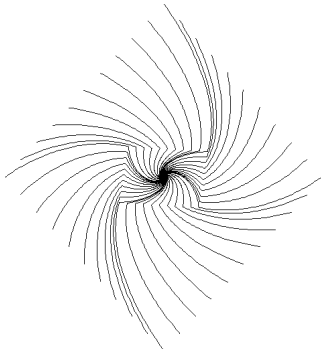
pour l'obtention de

L'HABILITATION À DIRIGER DES RECHERCHES

par

J.R. LOBRY

LE TROU NOIR DE L'ÉVOLUTION MOLÉCULAIRE SYMÉTRIQUE



Soutenue le 20 juillet 2000

Jury :

ANDERSSON, S.G.E.	(rapporteur)
DANCHIN, A.	(rapporteur)
GAUTIER, C.	(président)
LOUARN, J.-M.	
SUEOKA, N.	(rapporteur)

CNRS UMR 5558 « Biométrie, Biologie Évolutive » Université Claude Bernard - Lyon 1

INTRODUCTION

Il sera question ici de la théorie des pressions de mutation directionelles[164], ma contribution originale étant la mise en place d'un cadre théorique permettant de mettre en évidence des pressions de mutation directionnelles asymétriques, c'est à dire qui n'influencent pas de la même manière les deux brins de l'ADN. La notion de *direction* accolée à celle de *mutation* peut choquer en ce qu'elle pourrait suggérer un sous entendu de finalisme. Je vais montrer ci-après pourquoi il n'en est absolument pas question et préciser ce que l'on entend par pression de mutation directionnelle.

Finalisme et évolution

Tout finalisme en évolution a été abandonné depuis Darwin, ce que l'on peut schématiser en disant que l'évolution se fait selon un processus de Markov. Appelons information génétique un élément d'un ensemble E de k éléments discernables. Numérotons ces éléments pour pouvoir les désigner plus facilement et assimiler E aux k premiers entiers.

E = { 1, 2, 3, ..., k }

Une population génétique est un ensemble F de n informations génétiques. Soit $t_1, t_2, \quad , t_m$ une suite croissante d'instants et $X_{t_1}, X_{t_2}, \quad , X_{t_m}$ une suite de variables aléatoires à valeur dans E. La loi initiale des états est donnée par un vecteur de probabilité colonne,

$$\mathbf{P}_{t_0} = \begin{pmatrix} P(X_{t_0} = 1) \\ P(X_{t_0} = 2) \\ \dots \\ P(X_{t_0} = k) \end{pmatrix},$$

c'est-à-dire les fréquences relatives initiales des informations génétiques dans la population F. Soit **S** la matrice carrée d'ordre *k* des probabilités de passage dont le terme générique s_{ij} représente la probabilité d'avoir j à la date t_{m+1} sachant qu'il y avait un i à la date t_m,

$$s_{ij} = P\left(X_{t_{m+1}} = j \middle| X_{t_m} = i\right)$$

La loi des états à la date t_{m+1} est définie par P_{t_m} et la matrice des probabilités de passage S.

$$P_{t_{m+1}} = \mathbf{S} P_{t_m}$$

Dans un tel processus de Markov le passé n'influence l'avenir que via l'instant présent, il n'y a pas de finalisme possible. La théorie des pressions de mutation directionnelles satisfait cette condition, mais pour le voir il nous faut décomposer le passage de t_m à t_{m+1} en deux sous-étapes.

Le hasard et la nécessité

Les apports de la génétique et de la biologie moléculaire ont conduit à dissocier les processus générateurs de diversité, les mutations, qui se font au hasard, **H**, des processus de sélection naturelle de cette diversité **N**. Les mutations modifient les informations génétiques, c'est-à-dire l'aspect logiciel du vivant, tandis que la sélection ne voit que le coté matériel du vivant, la distinction entre les deux est imposée par l'irréversibilité de la transmission du flux d'information *in vivo* (ADN → ARN → Protéines), résultat fondamental de la biologie moléculaire.

$$P_{t_{m+1}} = \mathbf{N}\mathbf{H} P_{t_m}$$

L'évolution est un alors un processus de Markov simple alterné :

$$P_{t_0} \xrightarrow{\mathbf{H}} P_{t_{0bis}} \xrightarrow{\mathbf{N}} P_{t_1} \xrightarrow{\mathbf{H}} P_{t_{1bis}} \xrightarrow{\mathbf{N}} P_{t_2} \xrightarrow{\mathbf{H}} P_{t_{2\,bis}} \xrightarrow{\mathbf{N}} P_{t_3} \xrightarrow{\mathbf{H}} \ldots$$

Dans un tel processus le hasard **H** ne peut pas s'ajuster besoins de la nécessité **N** parce qu'il ne connaît que la population à l'instant présent, il ne peut pas tirer des leçons du passé pour anticiper le futur. C'est en ce sens que l'on dit que les mutations se font au hasard, comme l'ont très bien expliqué Graur et Li :

> ***« Are mutations random?***
>
> Mutations are commonly said to occur « randomly ». However, as we have seen mutations do not occur at random with respect to genomic location, nor do all types of mutation occur with equal frequency. So, what aspect of mutation is random? Mutations are claimed to be random in respect to their effect on the fitness of the organism carrying them. That is, any given mutation is expected to occur with the same frequency under conditions in which this mutation confers an advantage on the organism carrying it, as under conditions in which this mutation confers no advantage or is deleterious. »
>
> Graur and Li (2000) Fundamentals of molecular evolution[62].

On dit qu'il y a *mutation directionnelle* quand les probabilités de changements ne sont pas équiprobables, c'est-à-dire quand dans **H** les éléments hors diagonale ne sont pas tous égaux. Les mutations se font au hasard dans le sens de l'indépendance entre **H** et **N**, ce qui ne signifie pas l'équiprobabilité.

Dérive génétique et pression de mutation et de sélection

Neutralisme ne veut pas dire absence de sélection mais équiprobabilité de sélection des informations génétiques.

Hypothèse neutraliste :	$n_{ij} = \frac{1}{k}$

Comme cette matrice particulière ne modifie pas l'état du système, dans le cas du modèle markovien précédent l'évolution est contrôlée simplement par les taux de mutations,

$$P_{t_{m+1}} = \mathbf{H} P_{t_m},$$

cas particulier intéressant en tant qu'hypothèse nulle : c'est la théorie des pressions de mutations directionnelles énoncée par Sueoka en 1962[164].

En utilisant les fréquences *relatives* des informations génétiques pour caractériser une population nous avons fait l'hypothèse que l'effectif de la population est grand et constant au cours du temps. Dans la réalité les populations sont de taille finie : il y a des fluctuations d'échantillonnage d'une génération à l'autre conduisant à la dérive génétique.

> As an historical sidelight, Sueoka's theory appears to be one of the first neutral theories of DNA evolution. It explicitly assumes that natural selection plays no role in the dynamics of allele frequencies. However, as it does not incorporate genetic drift, it cannot describe the fixation of nucleotides. This aspect of the theory had to wait six more years for the publication of the papers by Kimura[96] and King and Jukes[98].
>
> John H. Gillespie. The causes of molecular Evolution[56].

Les choses ne sont pas si simples, le problème est que le concept de fixation d'une information génétique dans une population finie n'a de sens que pour des processus à caractère irréversible, c'est-à-dire quand k est suffisamment grand pour que l'on puisse dire qu'une nouvelle information génétique ne préexiste pas dans la population (modèle des allèles

infinis, des sites infinis, des mutations irréversibles) ou quand $n\mu << 1$, c'est dire quand le nombre de mutation par génération dans toute la population est faible.

Comme une information génétique n'a qu'une seule fréquence à un instant donné dans une population, la fonction de densité de probabilité à l'équilibre, $\varphi(x)$, ne représente pas quelque chose d'observable. Il faut imaginer un ensemble de populations évoluant dans les mêmes conditions, $\varphi(x)dx$ représente alors la fréquence relative des populations dont la fréquence relative d'une information génétique est dans l'intervalle $[x, x+dx]$.

Dans le cas d'une pression de mutation réversible avec k = 2 et en notant *u* et *v* les taux de mutation de et vers l'information génétique de fréquence x, Wright[186] a montré que la fonction de densité de probabilité est donnée par :

$$\varphi(x) = \frac{\Gamma(2n(u+v))}{\Gamma(2nu)\Gamma(2nv)} x^{2nv-1}(1-x)^{2nu-1}$$

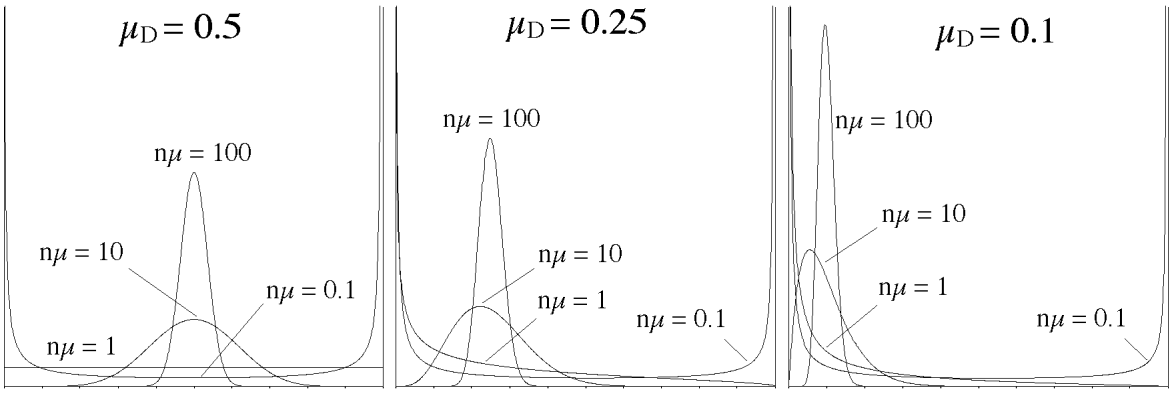
de moyenne,

$$\bar{x} = \int_0^1 x\varphi(x)dx = \frac{v}{u+v} = \mu_D$$

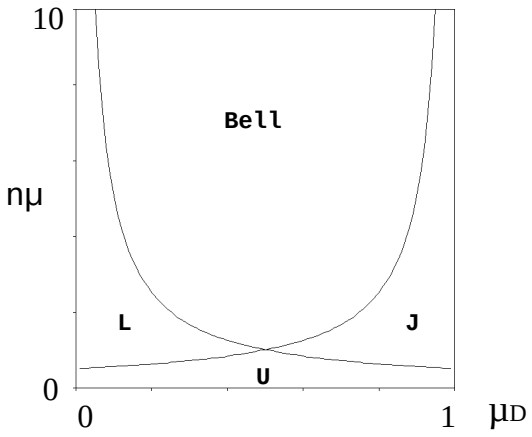
et de variance

$$\sigma_x^2 = \int_0^1 (x-\bar{x})^2 \varphi(x)dx = \frac{\mu_D(1-\mu_D)}{2n(u+v)+1}$$

La pression de mutation joue sur la moyenne et la variance de la distribution tandis que les effets des populations finies ne jouent que sur la variance. La figure ci-dessous représente $\varphi(x)$ dans différents cas de figure



L'allure de $\varphi(x)$ peut être celle d'une distribution en cloche, uniforme, en L ou en J selon la valeur du produit $n\mu$ et de μ_D,



la valeur critique $n\mu = 1$ signifie qu'il y a en moyenne exactement une mutation dans la population par génération, et la valeur critique $\mu_D = 0.5$ à l'égalité des taux de mutations *u* et *v*. On peut distinguer trois cas :

➔ Quand $n\mu >> 1$ et $\mu_D \approx 0.5$, c'est-à-dire quand il y a beaucoup de mutations à chaque génération dans la population et que la pression de mutation directionnelle n'est pas trop biaisée vers des fréquences extrêmes, on a un polymorphisme permanent dans la population. Le mode et la moyenne de la distribution sont proches, le plus probable est d'observer une population hétérogène avec x proche de μ_D. Il n'y a pas d'envahissement de la population par l'une ou l'autre information génétique, la notion de fixation n'a pas de sens ici.

➔ Quand $n\mu << 1$, c'est-à-dire quand il y a très peu de mutations à chaque génération, on retrouve les effets classiques de la dérive génétique. La population est le plus souvent homogène avec des transitoires polymorphiques permettant d'alterner les deux états monomorphiques de la population dont les probabilités sont voisines μ_D et de $1 - \mu_D$. La notion de fixation d'une information génétique a un sens ici.

➔ Quand $\mu_D \to 0.0$ ou $\mu_D \to 1.0$, c'est-à-dire quand la population est soumise à une forte pression de mutation dirigée vers des fréquences extrêmes, ou bien quand $n\mu \approx 1$, c'est-à-dire quand il y a en moyenne une mutation par génération dans la population, nous avons une situation un peu intermédiaire aux deux précédentes. Il y a un polymorphisme permanent dans la population mais le mode et la moyenne de la distribution sont très différents, le plus probable est d'observer une population monomorphique portant uniquement l'information génétique favorisée par le biais mutationel. La notion de fixation d'une information génétique n'a pas de sens ici, c'est toujours la même qui se fixe, nous sommes proches du domaine des pressions de mutations irréversibles[97].

Dans tous les cas, les effets de dérive génétique ne jouent que sur la variance de la distribution ce qui justifie l'intérêt de l'étude du modèle markovien où l'on s'intéresse à l'évolution des fréquences relatives moyennes des informations génétiques. Il faut simplement garder à l'esprit que cette moyenne est accompagnée d'une variance qui peut être très grande pour des petites populations.

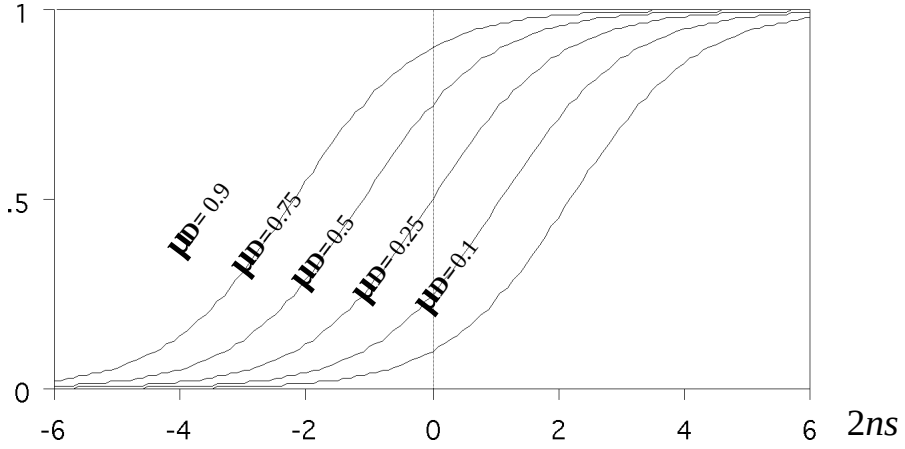
Dès que l'on sort de l'hypothèse neutraliste les modèles deviennent beaucoup plus complexes parce que les effets de dérive génétique vont également jouer sur la moyenne des fréquences relatives des informations génétiques. Toujours dans le cas $k = 2$ et avec un avantage sélectif de 0 pour l'information génétique de fréquence *x* et de *-s* pour celle de 1-*x*, Wen-Hsiung Li a montré[103] en 1987 que l'on avait approximativement :

$$\bar{x} \approx \frac{e^{2ns}v}{e^{2ns}v+u} = \mu_D \frac{e^{2ns}}{\mu_D e^{2ns}+1-\mu_D}$$

soit une courbe de réponse de type sigmoïde partant de μ_D à l'origine, quand il n'y a pas de sélection, pour tendre vers 1 pour les grandes valeurs du produit *ns*, quand la sélection est efficace. Le point de transition entre ces deux états est donné par l'abscisse *s** du point d'inflexion de la courbe,

$$s^* = \frac{1}{2n}\ln\left(\frac{1-\mu_D}{\mu_D}\right),$$

qui est fortement dépendant de la valeur de μ_D.



Quand $\mu_D = 0.5$ le point d'inflexion est à l'origine et l'on retrouve la condition classique $2ns >> 1$ pour que la sélection soit efficace, quand $\mu_D > 0.5$ le point d'inflexion a une valeur négative, les pressions de mutation et de sélection jouent dans le même sens, quand $\mu_D < 0.5$ le point d'inflexion a une valeur positive et les pressions de mutation et de sélection jouent en sens opposé.

> A mutation generally occurs in a single individual and give rise to an allele. If an allele achieves some frequency in a population it can be referred to as a polymorphism (not a « common [or rare] mutation ») If it has become fixed in a population it may be referred to as a substitution.
> *Molecular Biology and Evolution*, Instructions to Authors[7].

Dans la suite j'appelle taux de substitution r_{ij} la probabilité de passage de *i* à *j* par unité de temps,

$$r_{ij} = \frac{s_{ij}}{\Delta t} = \frac{P\left(X_{t_{m+1}} = j \middle| X_{t_m} = i\right)}{t_{m+1} - t_m},$$

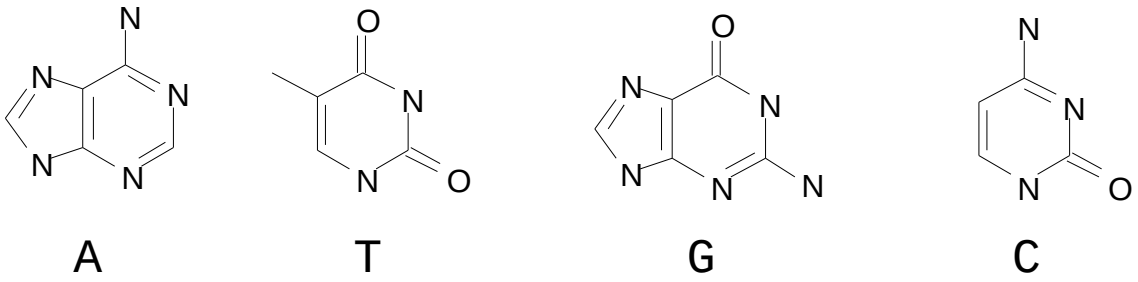
ce qui représente donc le résultat net instantané de la mutation et de la sélection. Il a une signification plus large que ce que l'on entend habituellement par taux de substitution de gène allélomorphe parce que l'on veut pouvoir traiter également les cas de polymorphisme permanent où il n'y a pas de fixation d'une information génétique dans la population.

Le support matériel des informations génétiques

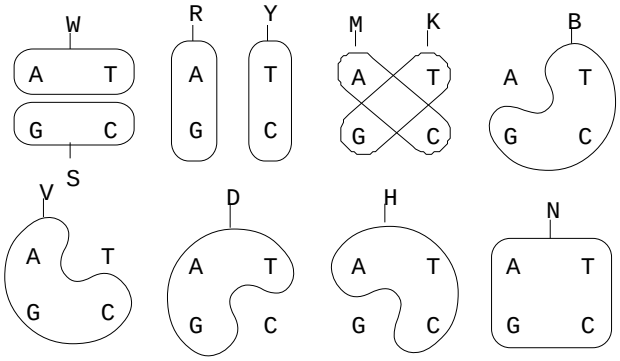
La génétique des populations est une discipline abstraite parce qu'elle est totalement indépendante du support matériel des informations génétiques : ses résultats seraient parfaitement valides pour des populations utilisant autre chose que les acides nucléiques.

Pourquoi devrait-on se soucier de ce support matériel ? C'est une question de niveau d'analyse, pour reprendre la métaphore de la barque de Delphes[35,36], quand on travaille au niveau des planches remplacer une planche en bois par une planche de sable ne serait pas sans conséquences sur les propriétés globales de la barque, et de même quand on travaille au niveau moléculaire il est difficile de faire complètement abstraction des propriétés physico-chimiques des bases.

Le support matériel des informations génétiques se présente sous la forme d'un hétéropolymer d'acides désoxyribonucléiques (ADN) de quatre types différant par leur base nucléique qui peut être une adénine (A), une thymine (T), une guanine (G), ou une cytosine (C).



Les abréviations standard[6] représentées ci-dessous seront employées dans la suite.



Dans de l'ADN bicaténaire les bases en vis-à-vis sont toujours W ou S[184], elles sont dites complémentaires. Notons $\overline{X}$ la base complémentaire de X, nous avons alors :

$$\overline{A} = T,\ \overline{T} = A,\ \overline{G} = C,\ \overline{C} = G$$

Autrement dit, quelle que soit la base N, nous avons toujours

$$\overline{\overline{N}} = N$$

C'est cette propriété du support matériel des informations génétiques que nous utiliserons dans la suite pour construire le modèle.

LE MODÈLE D'ÉVOLUTION SYMÉTRIQUE

Les hypothèses biologiques

L'hypothèse biologique fondamentale du modèle d'évolution symétrique est que les deux brins de l'ADN se comportent rigoureusement de la même façon aussi bien pour la mutation que pour la sélection. Cette hypothèse a été appelée règle de parité numéro 1 par Sueoka[169], soit RP1 en abrégé. Quelle en sont les conséquences au niveau de la structure de la matrice des taux de substitution ?

Notons

$r(X \to Y)$

le taux de substitution de X vers Y sur un brin, et

$\bar{r}(\bar{X} \to \bar{Y})$

le taux de substitution de l'événement complémentaire sur le brin complémentaire. Ces deux événements conduisant au même résultat le taux de substitution apparent sur un brin, $R(X \to Y)$, est égal à la somme de ces deux taux intrinsèques.

$$R(X \to Y) = r(X \to Y) + \bar{r}(\bar{X} \to \bar{Y})$$

On a de même pour la substitution complémentaire

$$R(\bar{X} \to \bar{Y}) = r(\bar{X} \to \bar{Y}) + \bar{r}(\bar{\bar{X}} \to \bar{\bar{Y}})$$

et comme

$\bar{\bar{N}} = N$

on a

$$R(\bar{X} \to \bar{Y}) = r(\bar{X} \to \bar{Y}) + \bar{r}(X \to Y)$$

L'hypothèse RP1 est que les taux de substitution sont les mêmes pour les deux brins,

Hypothèse RP1: $\quad X,Y \quad N : \; r(X \to Y) = \bar{r}(X \to Y)$

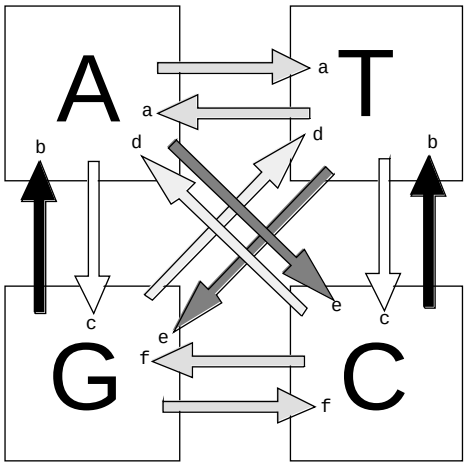
Par conséquent au niveau des taux apparents on aura

$$R(X \to Y) = R(\bar{X} \to \bar{Y})$$

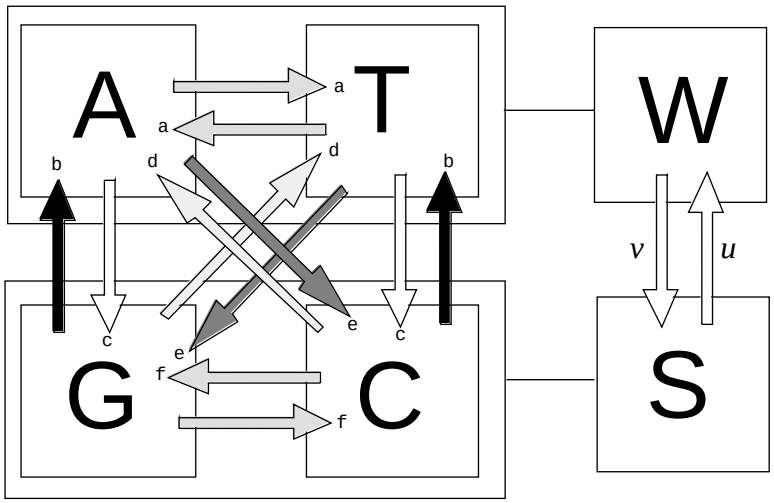
Ainsi, sous RP1, le taux de substitution apparent d'une base vers une autre est égal au taux de substitution de l'événement complémentaire, par exemple

$$R(A \to G) = R(T \to C)$$

Le nombre de taux de substitution est de 12 dans le modèle général, il est diminué de moitié grâce à l'hypothèse de symétrie entre les deux brins, comme le représente la figure ci-dessous :



Le modèle des taux de substitution sous RP1 peut être compris comme une simplification du modèle général à 12 paramètres, ou bien comme une extension du modèle de Sueoka[164] à deux paramètres,



la relation entre les deux modèles est donnée par $u = b + d$ et $v = e + c$, les deux paramètres a et f n'apparaissant plus de par la fusion des bases de types W et S dans le modèle à deux paramètres. Les transitions sont des substitutions intra R ou Y et correspondent aux paramètres b et c, les transversions isotypiques sont intra W ou S et correspondent aux paramètres a et f, les transversions allotypiques sont intra M ou K et correspondent aux paramètres d et e.

Notation du modèle

Soit **X** la matrice colonne,

$$\mathbf{X} = \begin{pmatrix} A(t) \\ T(t) \\ G(t) \\ C(t) \end{pmatrix}$$

dont les éléments sont les fréquences des nucléotides à l'instant t. Soit **R** la matrice du processus de l'évolution des fréquences des bases.

$$\frac{d\mathbf{X}}{dt} = \mathbf{R}\mathbf{X}$$

Les éléments de la matrice **R** sont les taux de substitution. De nombreuses formes paramétriques de la matrice **R** ont été publiées[104,106,149,189], celle qui nous intéresse ici est donnée par :

$$\mathbf{R} = \begin{pmatrix} -a-e-c & a & b & d \\ a & -a-e-c & d & b \\ c & e & -b-d-f & f \\ e & c & f & -b-d-f \end{pmatrix}$$

où les six paramètres (a,b,c,d,e,f) représentent les six taux de substitution comme dans la figure précédente. Les notations sont celles adoptées par Sueoka[169] et Lobry[107] en 1995. Ce modèle a été justifié et étudié indépendamment par Valenzuela[180] en 1997, il a également été posé par Wu et Maeda[188] en 1987 mais sans justification biologique.

Fréquences des bases à l'équilibre (RP2)

La position d'équilibre $\mathbf{X}^*$ est donnée par,

$$\mathbf{X}^* = \frac{1}{2}\begin{pmatrix} 1-\theta^* \\ 1-\theta^* \\ \theta^* \\ \theta^* \end{pmatrix}$$

où θ^* représente le taux de G+C à l'équilibre qui est ici uniquement déterminé[107] par 4 des 6 taux de substitution :

$$\theta^* = \frac{e+c}{b+c+d+e}$$

Ce résultat est cohérent avec le modèle de Sueoka à deux paramètres où le taux de G+C à l'équilibre est donné[52,164] par

$$\theta^* = \frac{v}{u+v}$$

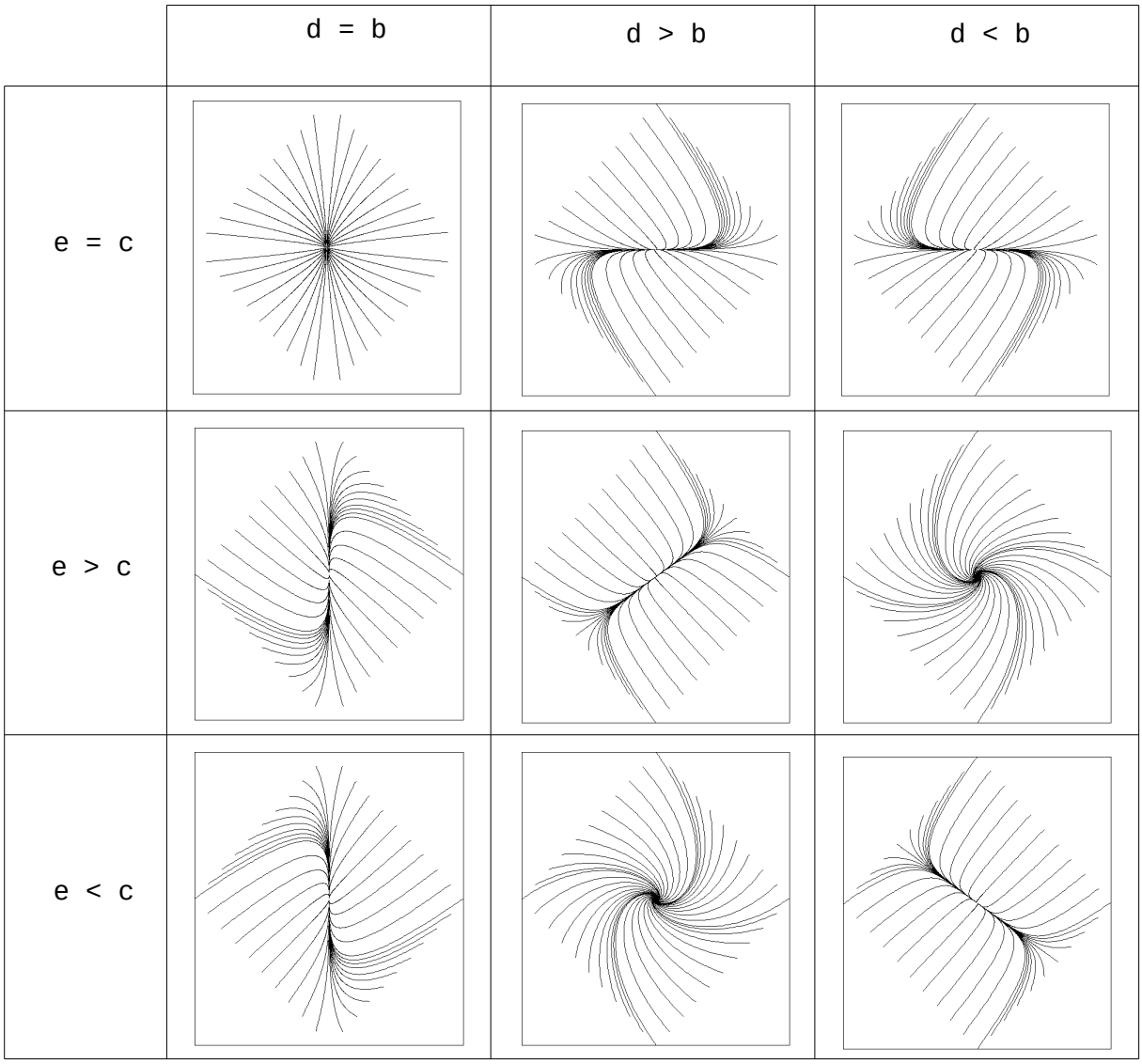
Ce point d'équilibre est tel que $A(t) = T(t)$ et $G(t) = C(t)$, c'est ce que Sueoka[169] a appelé la règle de parité numéro 2, soit RP2 en abrégé.

État RP2 : $A(t) = T(t)$ et $G(t) = C(t)$

Cette propriété fondamentale du modèle a été vérifiée indépendamment par Sueoka[169] avec des simulations numériques et par Valenzuela[180] analytiquement.

Convergence vers les fréquences de base à l'équilibre

Si on fait l'hypothèse que tous les taux de substitutions sont strictement positifs, alors la matrice **R** est une matrice compartimentale, et par conséquent elle n'a pas de valeur propre à partie réelle positive ou imaginaire pure[75]. De plus comme **R** correspond à un système clos sans piège interne, la multiplicité de la première valeur propre est égale à 1 de par le théorème de Foster et Jacquez[41]. Il n'y a donc qu'un seul point d'équilibre et ce point d'équilibre est stable : quels que soient les conditions initiales et les taux de substitution, on tend vers A=T et C=G[107], ce comportement est illustré ci-dessous pour différents cas de figure quant à la valeur des paramètres, l'axe des abscisses est A-T, celui des ordonnées est C-G, l'origine correspond à RP2.



La valeur des paramètres contrôle la façon dont on converge vers l'origine, on peut y aller de différentes façons, mais on y va de façon certaine. Si pour une séquence d'ADN on savait que l'équilibre est atteint alors on aurait un moyen simple de rejeter le modèle à partir de sa composition en bases. Malheureusement nous ne pouvons observer cette séquence qu'à

l'instant présent, toute déviation par rapport à RP2 peut alors s'interpréter alternativement comme un état hors équilibre sous RP1.

Convergence vers RP2

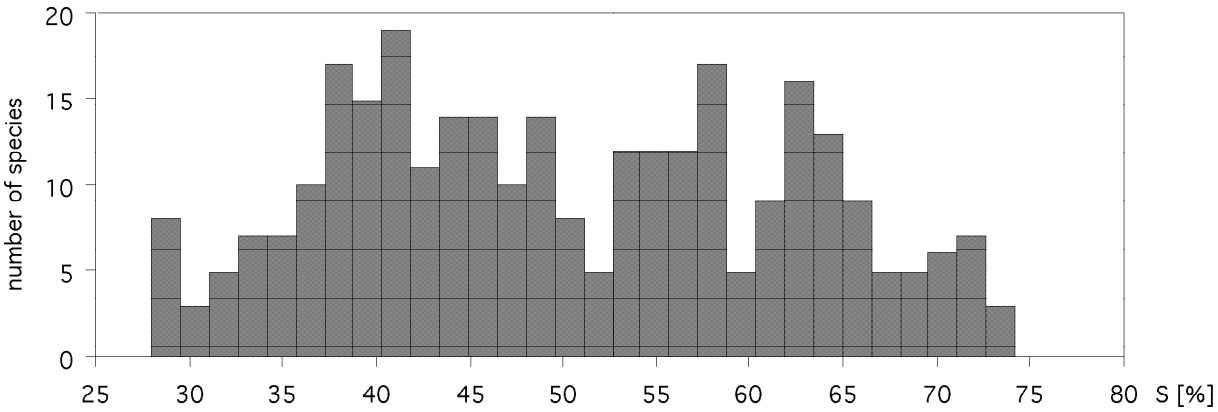
Nous avons montré[116] que l'on converge vers RP2 même dans le cas d'un système hors équilibre sous la condition que les taux de substitutions soient tous supérieurs à un seuil strictement positif donné. Ce résultat simple s'obtient avec un modèle plus compliqué dans lequel les taux de substitution sont autorisés a varier au cours du temps,

$$\frac{d\mathbf{X}}{dt} = \mathbf{R}(t)\mathbf{X},$$

mais où la matrice **R**(t) a toujours la structure imposée par RP1,

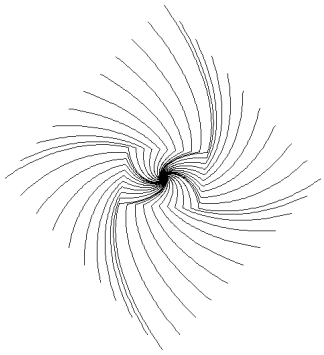
$$\mathbf{R}(t) = \begin{pmatrix} -a(t)-e(t)-c(t) & a(t) & b(t) & d(t) \\ a(t) & -a(t)-e(t)-c(t) & d(t) & b(t) \\ c(t) & e(t) & -b(t)-d(t)-f(t) & f(t) \\ e(t) & c(t) & f(t) & -b(t)-d(t)-f(t) \end{pmatrix},$$

puisque nous sommes toujours dans le cas d'une évolution parfaitement identique tant du point de vue de la mutation que de la sélection entre les deux brins d'ADN. D'un point de vue biologique ce modèle est plus réaliste car sur de grandes distances évolutives il n'est pas raisonnable de supposer que les taux de substitutions soient constants comme en témoigne la très grande variabilité de la fréquence des bases S dans les génomes bactériens[163]. La figure ci-dessous donne la distribution de la fréquence des bases S pour les 298 génomes bactériens où plus de 50 kb sont disponibles dans les banques.



Le trou noir de l'évolution moléculaire symétrique

Cette convergence vers RP2 est illustrée dans la simulation ci-dessous où les taux de substitution ont été modifiés en cours de route. Le système n'est alors plus à l'équilibre, en particulier pour la fréquence des bases S, mais cela ne l'empêche pas de continuer à converger vers RP2, même si c'est d'une manière différente.

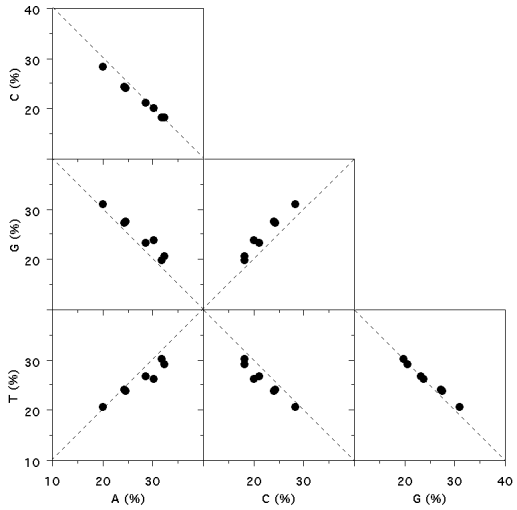


Ce résultat fait que nous sommes dans une situation beaucoup plus confortable pour pouvoir rejeter le modèle à partir des fréquences des bases de l'ADN puisque nous n'avons plus besoin de supposer que nous sommes à l'équilibre.

RP2 en tant qu'approximation pour les génomes complets

> « Not unrelated to this as yet unexplained finding may be later observations from my laboratory, namely, that in microbial DNA the separated heavy and light strands, although complementary to each other with respect to base composition, both exhibit the same equivalence of 6-amino and 6-oxo bases. To my knowledge, there have been no follow-up studies of the last-mentioned observations in other laboratories. »
>
> Erwin Chargaff (1979) How genetics got a chemical education[27].

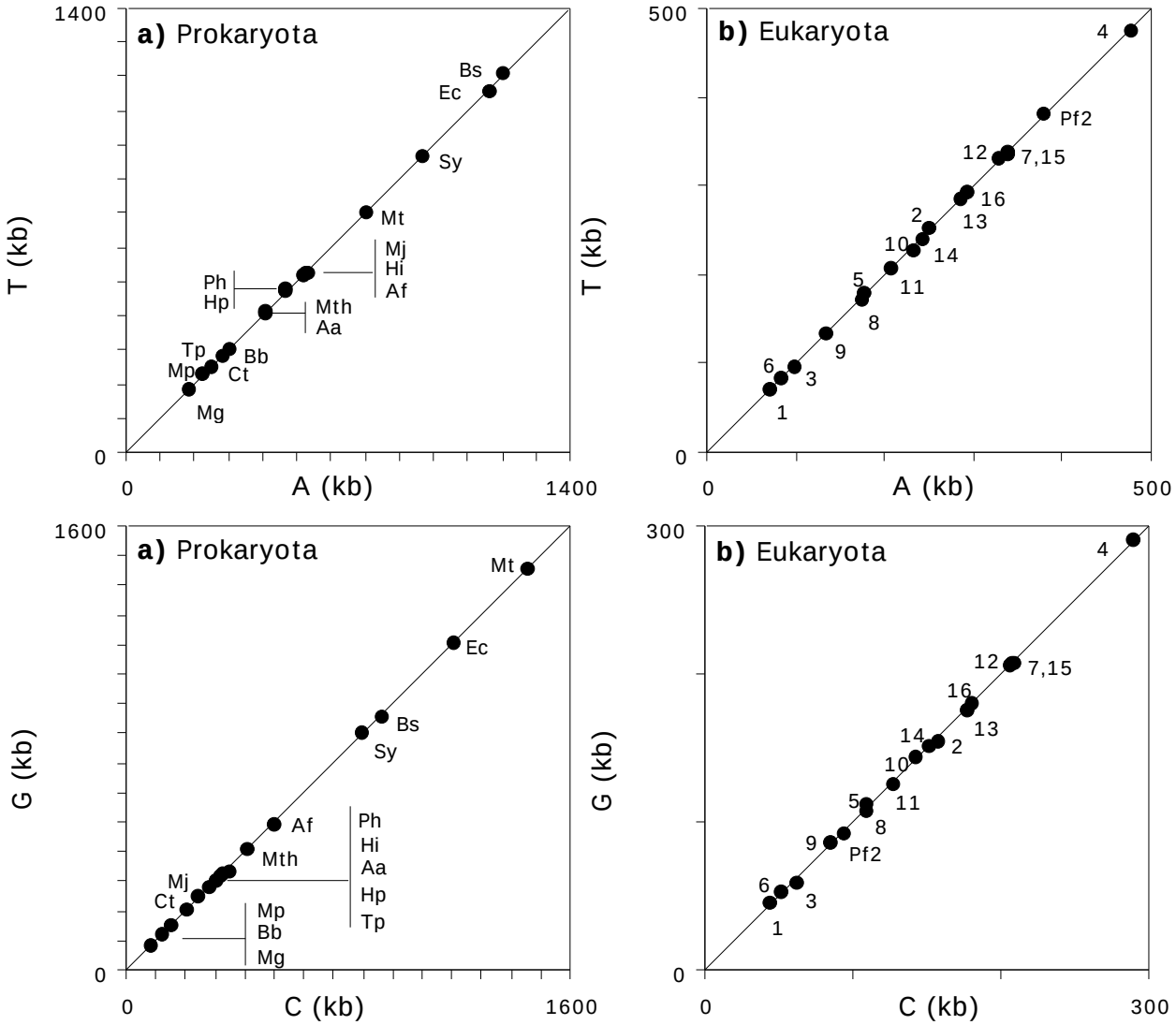
La mesure expérimentale de la composition globale d'un des deux brins d'un génome complet est délicate puisqu'il faut au préalable séparer les deux brins, sans quoi on retrouvera immanquablement RP2 à cause de la nature complémentaire de l'ADN bicaténaire. Cette analyse sur de l'ADN monocaténaire a été faite en 1968 avec le génome de *Bacillus subtilis* dans le laboratoire de Chargaff[152,90], puis généralisé avec le génome de 6 nouvelles espèces bactériennes[153] : *Proteus vulgaris*, *Bacillus megaterium*, *Bacillus stearothermophilus*, *Escherichia coli*, *Salmonella enterica* serovar Typhimurium et *Serratia marcescens*. La figure ci-dessous reprend ces résultats pour le brin L, les pointillés figurant ce que l'on devrait obtenir si on avait exactement RP2 dans de l'ADN monocaténaire.



Ces résultats étaient pour le moins troublant car si l'on comprend bien pourquoi on a RP2 dans de l'ADN bicaténaire, on ne voit pas pourquoi on aurait RP2 dans de l'ADN monocaténaire. Il est amusant de noter que si RP2 est vrai pour de l'ADN monocaténaire alors on ne peut plus utiliser le fait que l'on ait RP2 pour de l'ADN bicaténaire comme argument en faveur de la structure en double hélice de l'ADN[184]. Quoi qu'il en soit, ces résultats, bien que rappelés lors de l'étude des fréquences des oligonucléotides dans chaque brin[154], furent plus ou moins oubliés pendant un quart de siècle, avant que le séquençage de grands fragments génomiques permette de réexaminer cette question, avec une précision beaucoup plus fine pour ce qui est de l'estimation des fréquences des bases.

Nussinov fit remarquer en 1982 que pour trois génomes complets de virus eucaryotes on observe RP2 dans chaque brin[130], mais ces génomes sont très petits, de l'ordre de 5 kb. Une étude plus sytématique[38] avec toutes les séquences disponibles de *Homo sapiens* et *Escherichia coli* en 1992 et montra que l'on observe globalement RP2 pour des sous-séquences de taille comprise entre 50 bp et 1000 bp. Le problème de cette analyse est qu'elle mélange des séquences provenant de l'un ou l'autre brin de l'ADN, neutralisant ainsi une déviation éventuelle par rapport à RP2. L'étude de Prabhu[142] sur 32 longs fragments génomiques de plus de 50 kb montra que l'on a bien RP2 dans de l'ADN monocaténaire, résultat confirmé[107] quand 60 fragments de plus de 50 kb furent disponibles en juin 1994. Ces fragments provenant de génomes très divers (virus, procaryotes, nématode, chloroplastes, insectes, vertébrés, mitochondries, levure) l'observation semble assez générale.

Le génome de *Mycoplasma genitalium*[49] inaugure en 1995 l'ère des génomes complets autres que ceux d'organelles et de virus. L'analyse[116] des génomes complets de 4 archaebactéries, de 12 eubactéries, des 16 chromosomes de *Saccharomyces cerevisiae* et du chromosome 2 de *Plasmodium falciparum* montre que RP2 est une approximation assez satisfaisante pour de l'ADN monocaténaire :



Une interprétation sélectionniste de RP2 dans les génomes complets a été avancée par Forsdyke[40] : ce serait le résultat d'une pression de sélection favorisant les mutations engendrant des oligonucléotides complémentaires voisins, créant la possibilité de former des structures secondaires de type cruciforme. Cette interprétation n'est pas convaincante parce que i) elle n'est basée que sur la propension prédite[191] *in silico* des séquences d'ADN à former des structures secondaires alors qu'il n'est pas certain que les cruciformes existent *in vivo*[118,160]. ii) Les données expérimentales montrent que la stabilité des cruciformes diminue avec la température[133], s'ils existaient *in vivo* on s'attendrait à ce que la fréquence en bases S au niveau des tiges augmente avec la température[181], comme c'est bien le cas au niveau des tiges des ARNr et des ARNt, mais ce n'est pas ce qui est observé[55]. iii) La proportion de bases qui auraient une base complémentaire potentielle dans le même brin serait[11] $\left(W-|A-T|+S-|C-G|\right)/N$. Par exemple chez *Borrelia burgdorferi*[47] (A = 323079, T = 327196, C = 130760, G = 129646) il y aurait ainsi 99.4 % des bases impliquées dans de telles structures. Sachant que 93.6 % des bases sont dans des séquences codantes dans ce génome, on voit mal comment de telles contraintes sélectives pourraient être compatibles.

Ces résultats au niveau des génomes complets ne sont en fait pas très intéressants puisqu'ils ne nous conduisent pas à rejeter le modèle.

LA STRUCTURE EN CHIROCHORE DES GÉNOMES BACTÉRIENS

Signification du rejet du modèle

Les taux de substitution de la matrice **R** représentent le résultat de la mutation et de la sélection. Le rejet du modèle ne permet pas d'identifier la cause de l'asymétrie entre les deux brins, il faut pour cela des données biologiques supplémentaires pour pouvoir discuter de ce point intéressant.

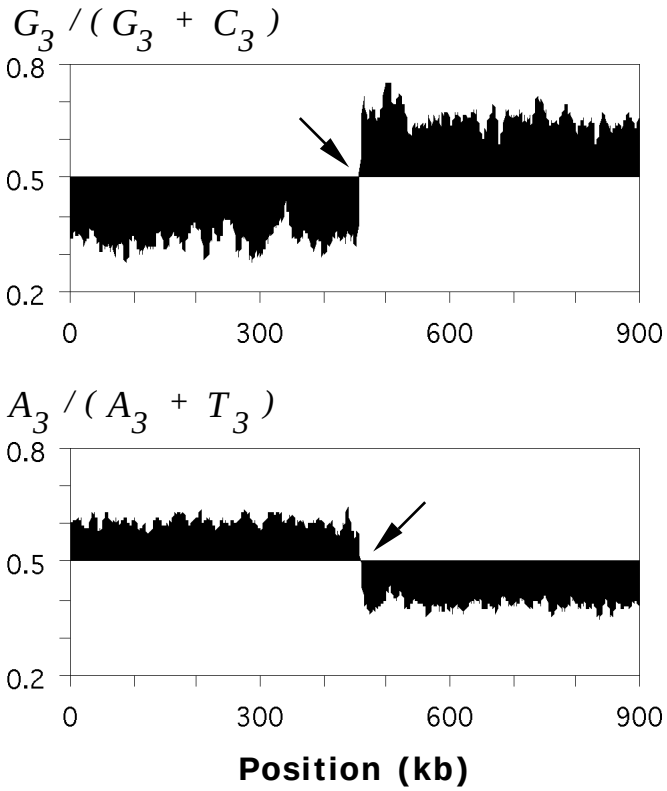
J'ai découvert[109] par hasard la structure en chirochores des génomes bactériens en voulant tester les prédictions du modèle d'évolution symétrique moléculaire. Chez les bactéries on observe souvent des segments homogènes du point de vue des déviations à RP2 que j'ai appelés chirochores, par analogie avec les isochores qui sont des segments homogènes du point de vue de la composition en bases S. Les chirochores sont purement descriptifs. Les réplichores[18] sont les segments du génome compris entre une origine et un terminus de réplication. Les frontières des chirochores et des réplichores coïncident chez les bactéries[108,109,110,51,128,63,65,92,125,94,124,155,101,146,147,117,25,26,119,120,121].

Une structure en chirochore a également été rapportée lors de la publication des génomes complets de *Escherichia coli*[18], *Bacillus subtilis*[99], *Borrelia burgdorferi*[47], *Rickettsia prowazekii*[5], *Campylobacter jejuni*[135], *Treponema pallidum*[48], *Nesseria meningitidis*[176,134], *Chlamydia trachomatis*[144], *Chlamydia pneumoniae*[144].

La structure en chirochore des génomes bactériens est le résultat de superpositions complexes entre les pressions de sélection et de mutation asymétriques[46], l'analyse de la variance à deux facteurs[179] (brin sens contre anti-sens, brin précoce contre brin tardif) a montré qu'une proportion significative des biais de composition en bases chez les bactéries pouvait être attribués à l'orientation par rapport à la réplication : la composition d'un gène est différente selon qu'il est sur le brin précoce ou tardif.

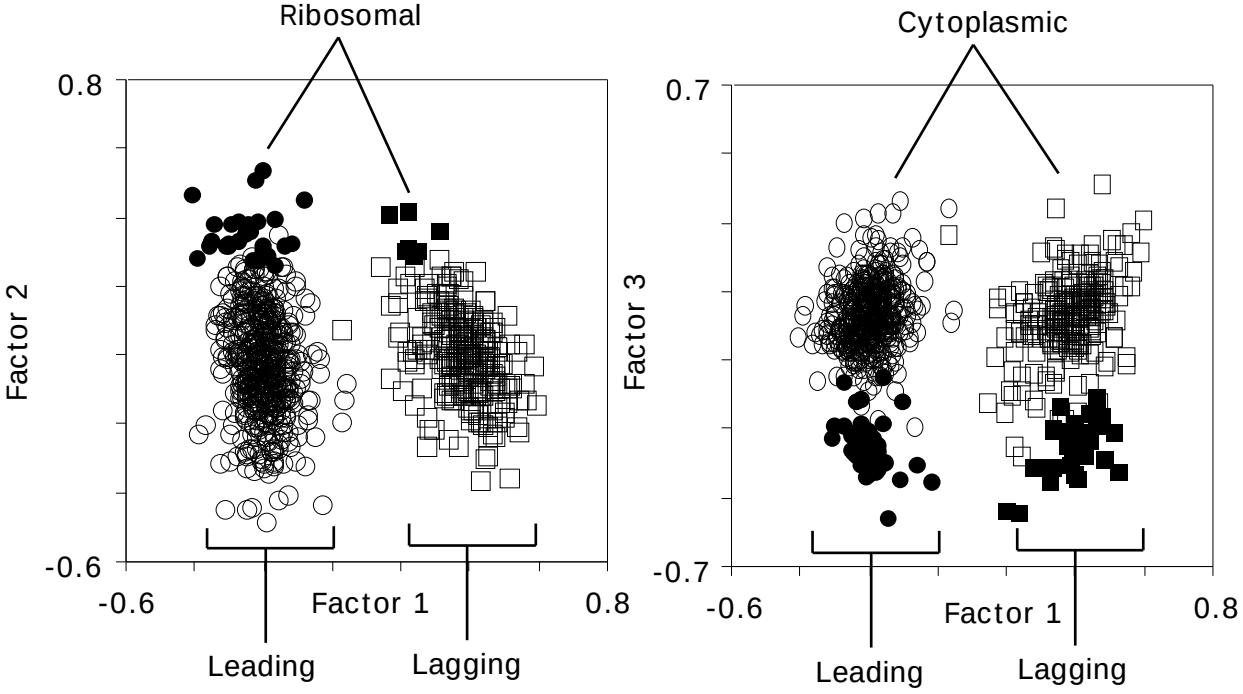
Exemple de pression de mutation asymétrique

La structure en chirochore la plus prononcée connue à ce jour est celle de *Borrelia burgdorferi*, représentée ci-dessous avec une fenêtre glissante de 10kb et un pas de 1kb en ne retenant que les bases en position 3 des codons sur le brin publié[47] dans les banques,



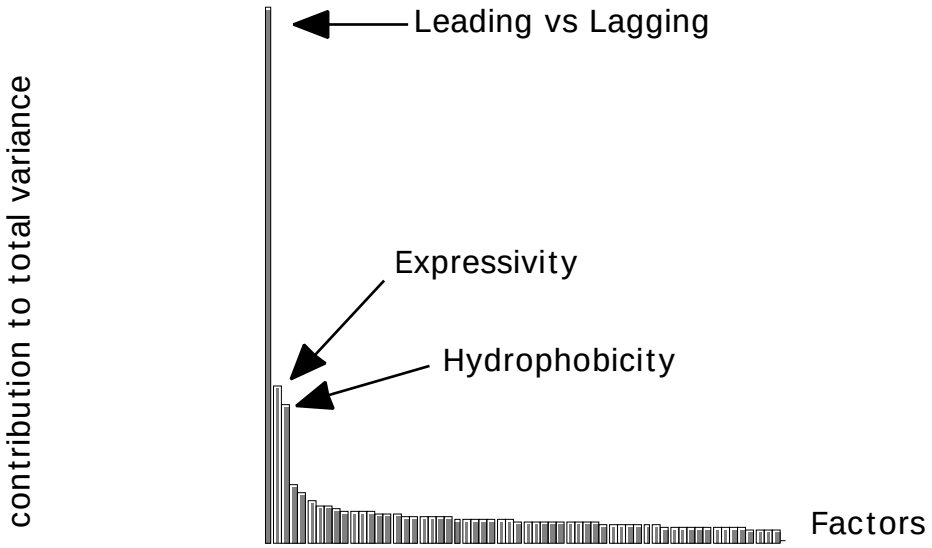
les flèches pointent vers l'origine de réplication du chromosome telle qu'elle a été déterminée expérimentalement[141].

La structure en chirochore se répercute sur l'usage du code de *B. burgdorferi* comme le montre les deux premiers plans factoriels de l'analyse factorielle des correspondances des fréquences des codons des 772 séquences codantes de plus de 300 bp de ce génome :



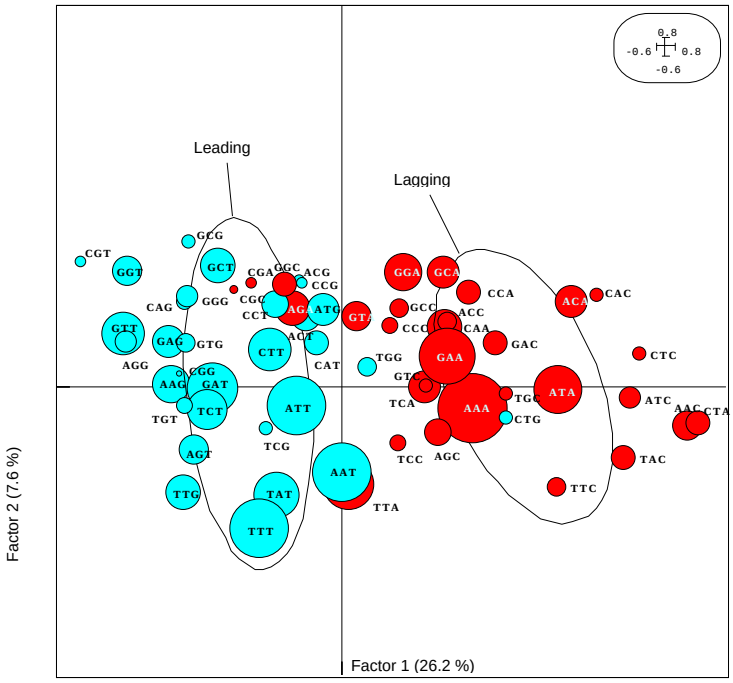
Le premier facteur (26.2 %) est l'opposition entre les séquences codantes localisées sur le brin précoce pour la réplication contre celles du brin tardif. Sur le deuxième facteur (7.6 %) on

retrouve l'effet classique[57,74] lié au degré d'expressivité des gènes, et sur le troisième facteur (6.7 %) l'opposition classique[115] entre les séquences codant pour des protéines membranaires intégrales et celles codant pour des protéines cytoplasmiques. On notera que les plans factoriels sont moins brouillés que ce qui a pu être publié par ailleurs[124,101] parce que j'ai utilisé ici les fréquences absolues des codons et non les RSCU[157]. Le graphe des valeurs propres permet de visualiser simplement la contribution relative des différents facteurs interprétables et résiduels.

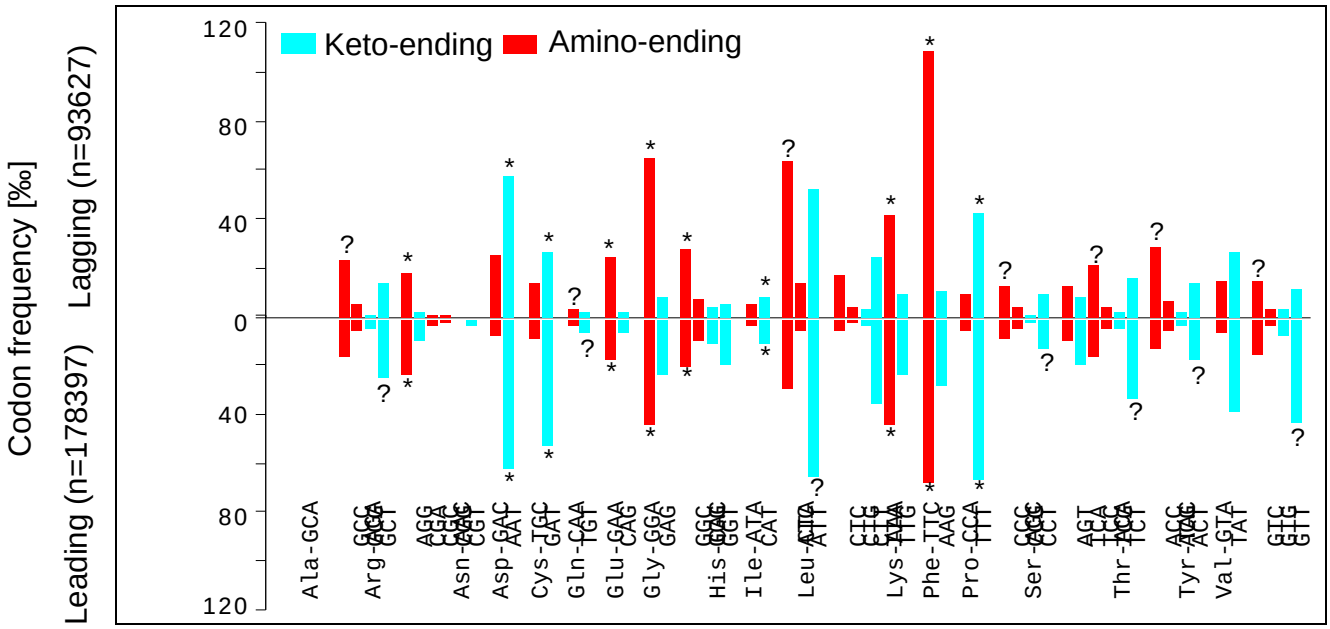


Le génome de *B. burgdorferi* est donc assez extraordinaire avec deux sous ensembles de séquences codantes ayant un usage du code radicalement différent dans un même génome. *B. burgdorferi* est dans un véritable cul-de-sac évolutif dur point de vue de l'optimisation de la traduction, il ne lui resterait plus qu'à transférer tous ses gènes sur le brin précoce, un peu comme chez certaines mitochondries, pour ne plus avoir à gérer deux usages du code différents.

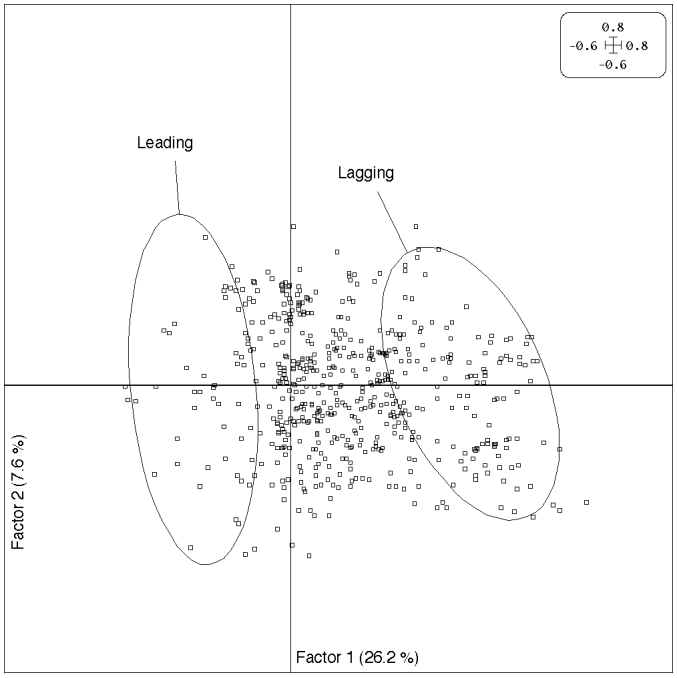
Le premier plan factoriel dans l'espace des codons ci-dessous montre que les deux groupes de séquences codantes diffèrent principalement en position 3 des codons avec les séquences du groupe précoce enrichies en bases K (gris clair) alors que celles du groupe tardif le sont bases M (gris foncé).



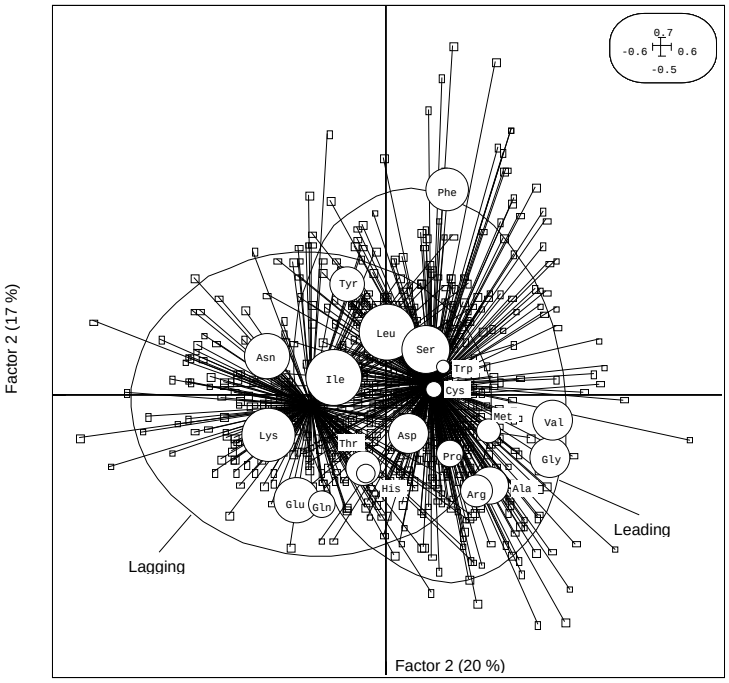
Les positions 3 des codons étant les moins contraintes sélectivement, c'est la mise en évidence de pression de mutations directionnelles asymétriques dans ce génome. Chez les organismes unicellulaires, il est bien connu que le premier facteur de la variabilité des fréquences des codons est lié à l'expressivité[74,57,70,158,156,3,83]: les codons les plus fréquents correspondent aux ARNt les plus abondants et ce d'autant plus que le niveau d'expression des gènes est plus important. Cette pression sélective est suffisamment importante pour moduler la composition des protéines chez *Escherichia coli*[159,115]. Chez *Borrelia burgdorferi*, la pression de mutation asymétrique et l'optimisation de la traduction jouent en sens opposé. Par exemple, parmi les codons AAR codant pour Lys c'est le codon AAA qui est majoritaire aussi bien pour les gènes sur le brin précoce que sur le brin tardif. Du point de vue de l'optimisation de la traduction, les gènes auraient intérêt à se trouver sur le brin tardif pour bénéficier de la pression de mutation asymétrique qui enrichit le brin tardif en A. En revanche, parmi les codons AAY codant pour Asn c'est le codon AAT qui est majoritaire pour les gènes des deux groupes, les gènes auraient intérêt cette fois à se trouver sur le brin précoce pour bénéficier de la pression de mutation asymétrique qui enrichit le brin précoce en T. On ne peut pas généraliser ce raisonnement à tous les acides aminés parce que la pression de mutation asymétrique est si forte que ce n'est plus le même codon qui est majoritaire dans les deux groupes (points d'interrogations dans le graphe ci-dessous). Mais les codons optimaux qui peuvent être déterminés (étoiles) sont favorisés par la pression de mutation asymétrique soit sur le brin précoce soit sur le brin tardif.



Les 12 plasmides linéaires et les 9 plasmides circulaires de *Borrelia burgdorferi* comportent plus de 40 % des gènes de la cellule, et il a été proposé que ces plasmides soient en fait des minichromosomes[10]. La projection en individus supplémentaires des gènes plasmidiques sur le premier plan factoriel est représentée ci-dessous. Elle met en évidence le fait que les gènes plasmidiques sont en général moins biaisés, ce qui pourrait s'expliquer par le fort taux de flux génétique, et donc d'inversion, chez ces plasmides[23].



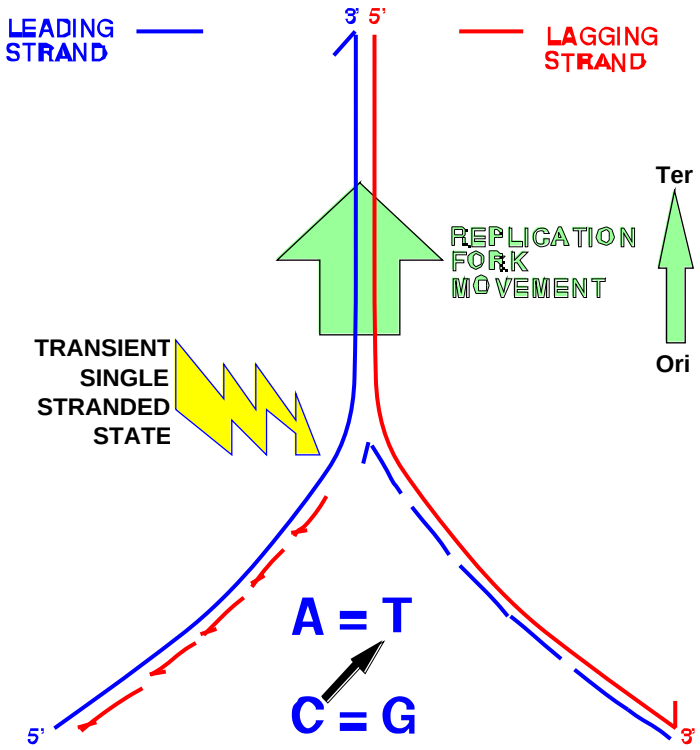
La pression de mutation asymétrique est suffisamment importante chez *Borrelia burgdorferi* pour moduler la composition en acides aminés des protéines[147,101,119]. L'analyse factorielle des correspondances de la composition en acides aminés des protéines représentée ci-dessous montre que le premier facteur de variabilité est l'orientation par rapport à la réplication, alors qu'habituellement c'est l'opposition entre les protéines cytoplasmiques et les protéines membranaire qui est le premier facteur de variabilité[115].



J'ai insisté ici sur le cas de *Borrelia burgdorferi* parce qu'il est le plus spectaculaire. Il faut souligner cependant que cette observation est plus générale. Par universalité du biais on entend que s'il est détectable il est toujours orienté dans le même sens avec le brin précoce enrichi en bases K[146], et non que ce biais existe toujours[92]. Chez les bactéries, l'analyse discriminante des correspondances a montré[140] que le biais existait dans les génomes de *Escherichia coli*, *Haemophilus influenzae*, *Bacillus subtilis* et *Mycoplasma genitalium*. Cette observation a été étendue[147] au cas des génomes de *Borrelia burgdorferi*, *Chlamydia trachomatis*, *Helicobacter pylori*, *Methanobacterium thermoautotrophicum*, *Mycobacterium tuberculosis*, et *Treponema pallidum*. Sur 22 génomes bactériens complets, le biais est détectable pour 16 génomes[114].

En dehors des eubactéries et des archéobactéries, il est intéressant de noter que ce biais universel a été détecté dans le génome du chloroplaste d'*Euglena gracilis*[127], chez des virus[39,63,64,128], et des mitochondries[8,175,76,137,145,143]. La controverse sur l'existence d'une pression de mutation asymétrique dans la région du gène de la ß-globine chez l'homme[188,22,187] semble maintenant terminée[45] : rien de significatif ne semble pouvoir être mis en évidence.

Proposée initialement pour les mitochondries[21], la théorie de la désamination des cytosines est fondée sur le fait que cette réaction est 140 fois plus rapide dans de l'ADN monocaténaire que dans de l'ADN bicaténaire[50]. Pendant la réplication, le brin tardif est protégé par le brin précoce néosynthétisé alors que le brin précoce reste transitoirement à l'état simple brin en attendant que le brin tardif néosynthétisé soit assez long pour restaurer l'état bicaténaire[9,122,129].



Cette asymétrie fondamentale de la réplication pourrait expliquer l'universalité des biais observés. Les arguments les plus forts sont trouvés chez des mitochondries[145] et des virus[63,64] où le temps d'exposition à l'état simple brin est corrélé avec l'intensité des biais. La protection contre la désamination des cytosines pourrait varier d'une espèce à l'autre et expliquer la variabilité de l'intensité des biais. La taille des fragments d'Okazaki étant beaucoup plus courts (0.1-0.2 Kb) chez les vertébrés que chez les bactéries (1-2 Kb) expliquerait l'absence de biais détectable chez les vertébrés[45]. Ce modèle simple explique également pourquoi les biais sont moins forts pour les bases W que pour les bases S. En effet, partons d'une séquence à l'état RP2, notons Θ_0 la fréquence relative initiale des bases S, et supposons *en première approximation* que l'effet de la désamination des cytosines soit de transformer une fraction α des bases C en T. On a alors pour l'excès de bases G,

$$\frac{G}{G+C}(\alpha) = \frac{\frac{1}{2}\Theta_0}{\frac{1}{2}\Theta_0 + \frac{1}{2}\Theta_0 - \frac{1}{2}\alpha\Theta_0} = \frac{1}{2-\alpha},$$

une expression indépendante du taux de bases S initial qui vaut au maximum 1 pour $\alpha = 1$. Par contre pour l'excès de bases T,

$$\frac{T}{A+T}(\alpha) = \frac{\frac{1}{2}(1-\Theta_0) + \frac{1}{2}\alpha\Theta_0}{\frac{1}{2}(1-\Theta_0) + \frac{1}{2}(1-\Theta_0) + \frac{1}{2}\alpha\Theta_0} = \frac{1-\Theta_0+\alpha\Theta_0}{2-2\Theta_0+\alpha\Theta_0},$$

nous avons une expression qui dépend de la fraction initiale des bases S et qui vaut au maximum $1/(2-\Theta_0)$ pour $\alpha = 1$. Ce n'est que dans le cas très particulier où il n'y avait pas de bases W au départ ($\Theta_0 = 1$) que l'on peut atteindre la valeur maximale de 1 comme pour l'excès de bases G.

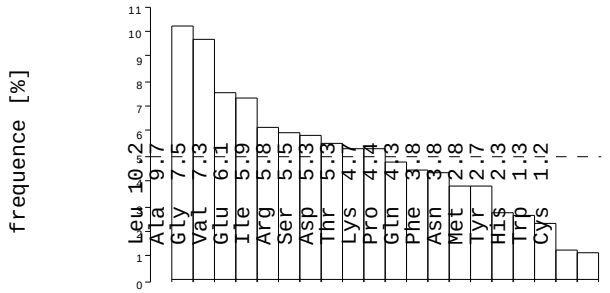
On s'attend donc à ce que l'excès de T soit plus faible que l'excès de G, et c'est bien ce qui est observé. Par exemple, en partant d'une séquence où toutes les bases ont la même

fréquence ($\Theta_0 = 0.5$) et que nous transformons tous les C en T (α = 1) l'excès de bases G, G/(G+C) = 1, est supérieur à l'excès de bases T, T/(A+T) = 2/3. La désamination des cytosines est donc compatible, au moins qualitativement, avec l'universalité de l'orientation des biais, ce qui ne signifie pas que ce soit le seul phénomène présent. La modélisation de l'effet de la désamination comme une simple transformation statique d'une fraction α des bases C en T est assez brutale, il serait préférable de modéliser directement au niveau de la matrice des taux de substitution.

Exemple de pressions de sélection asymétrique

Les fréquences moyennes des acides aminés des protéines sont modulées par les pressions de mutation symétriques chez les bactéries[162,111,66,185,33,28], les virus[16,93,12,20], les mitochondries[79,77] et les eucaryotes[16,166,67,34,167,29,33,32]. Il n'est d'ailleurs pas possible d'étudier les adaptations thermophiliques des protéines des micro-organismes sans tenir compte de cet effet[68,123]. Cependant, l'amplitude des variations observées est moindre que si les fréquences moyenne en acides aminés étaient libres de toutes contraintes sélectives[111]. Il est donc raisonnable d'interpréter ceci comme l'existence d'une pression de sélection centripète sur les fréquences des acides aminés autour d'une concentration optimale pour que les propriétés physico-chimiques globales des protéines, comme leur solubilité, ne soient pas trop aberantes[80].

Les fréquences optimales en acides aminés sont inconnues, mais on peut utiliser à titre indicatif celles de *Escherichia coli*[115], qui avec un taux de G+C génomique voisin de 50.8 % occupe une position médiane dans la gamme observée chez les bactéries[163,171].



Pour calculer le rapport attendu T_1/A_1 il nous suffit de calculer le rapport des fréquences des codons TNN/ANN compatibles avec les fréquences des acides aminés, comme dans le tableau ci-dessous :

TNN	aa	aa%	min	max	uni	ANN	aa	aa%	min	max	uni
TTY	Phe	3.8	3.8	3.8	3.8	ATH	Ile	5.9	5.9	5.9	5.9
TTR	Leu2	10.2	0.0	10.2	3.4	ATG	Met	2.8	2.8	2.8	2.8
TCN	Ser4	5.5	0.0	5.5	3.7	ACN	Thr	5.3	5.3	5.3	5.3
TAY	Tyr	2.7	2.7	2.7	2.7	AAY	Asn	3.8	3.8	3.8	3.8
TGY	Cys	1.2	1.2	1.2	1.2	AAR	Lys	4.7	4.7	4.7	4.7
TGG	Trp	1.3	1.3	1.3	1.3	AGY	Ser2	5.5	0.0	5.5	1.8
						AGR	Arg2	5.8	0.0	5.8	1.9
Sum			9.0	24.7	16.1			22.5	33.8	26.2	

T1/A1 max = 24.7/22.5 = 1.10
T1/A1 uni = 16.1/26.2 = 0.61
T1/A1 min = 9.0/33.8 = 0.27

On s'attend ainsi à avoir le rapport T_1/A_1 varier de 0.27 à 1.10 pour des usages du code extrêmes pour les acides aminés dont les codons ne commencent pas tous par T ou A (Leu, Ser et Arg), avec une valeur attendue de 0.61 si pour ces acides aminés les codons sont uniformément répartis entre les deux classes. Le déficit de T en première position des codons s'explique ainsi principalement parce que les codons TRN codent pour trois acides aminés des plus rares (Tyr, Trp, Cys) ainsi que pour les trois codons stop.

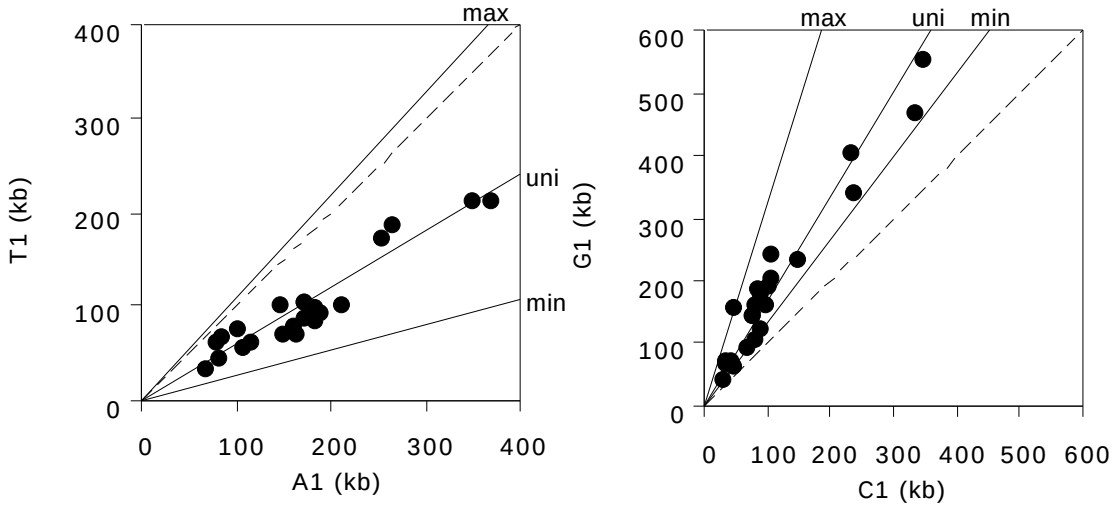
De manière analogue on peut calculer les rapports G_1/C_1 attendus,

GNN	aa	aa%	min	max	uni	CNN	aa	aa%	min	max	uni
GTN	Val	7.3	7.3	7.3	7.3	CTN	Leu4	10.2	0.0	10.2	6.8
GCN	Ala	9.7	9.7	9.7	9.7	CCN	Pro	4.4	4.4	4.4	4.4
GAY	Asp	5.3	5.3	5.3	5.3	CAY	His	2.3	2.3	2.3	2.3
GAR	Glu	6.1	6.1	6.1	6.1	CAR	Gln	4.3	4.3	4.3	4.3
GGN	Gly	7.5	7.5	7.5	7.5	CGN	Arg4	5.8	0.0	5.8	3.9
Sum			35.9	35.9	35.9				11.0	27.0	21.7

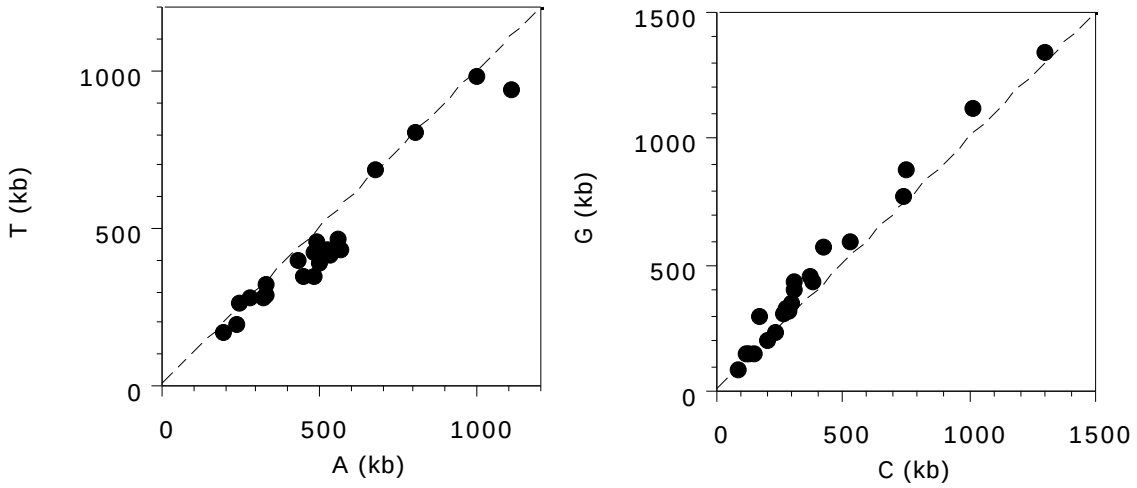
G1/C1 max = 35.9/11.0 = 3.26
G1/C1 uni = 35.9/21.7 = 1.65
G1/C1 min = 35.9/27.0 = 1.33

pour constater que l'excès de G en première position s'explique principalement parce que les codons GNN correspondent à des acides aminés fréquents dans les protéines.

Le graphe ci-dessous représente ce qui est observé dans les génomes complets bactériens disponibles, l'excès moyen de A par rapport à T (T_1/A_1 = 0.60), et l'excès de G par rapport à C (G_1/C_1 = 1.50) correspondent bien aux valeurs attendues. Un cas assez exceptionnel, déjà remarqué[183], est celui de *Methanococcus jannaschii* qui présente un ratio de G_1/C_1 de 3.8 exacerbé par la faible fréquence en His (1.4 %) et Gln (1.4 %) chez cet organisme.



Un raisonnement analogue permet de calculer les proportions attendues pour les autres positions des codons (T_2/A_2 = 1.03, G_2/C_2 = 0.76, T_3/A_3 = 1.01, G_3/C_3 = 1.02). L'excès de A en première position n'est pas compensé par les autres positions, l'excès de G en première position est partiellement compensé par la deuxième position, toutes positions confondues les rapports attendus sont T/A = 0.88 et G/C = 1.14, c'est-à-dire un léger excès de bases R dans les séquences codantes, ce qui correspond bien à ce qui est observé[174].



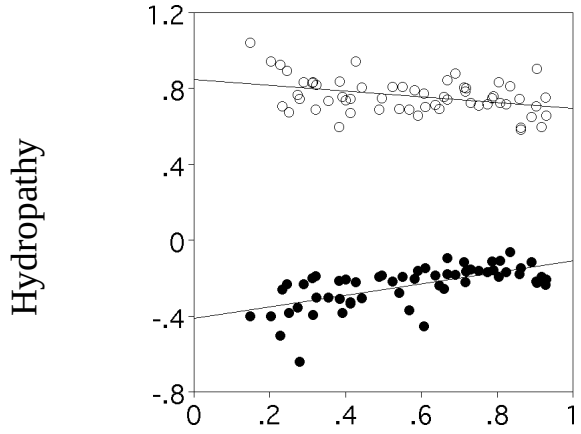
Ces biais sélectifs ne permettent pas à eux seuls de conduire à une structure en chirochore, car si les séquences codantes étaient réparties équitablement entre les deux brins, ils se neutraliseraient en moyenne. Cependant chez les bactéries il y a souvent un excès de gènes sur le brin précoce, ce qui est interprété comme une contre sélection pour éviter les collisions frontales entre la fourche de réplication et l'ARN polymerase[19,190,105]. Sous cette hypothèse, la pression de sélection pour orienter les gènes dans le bon sens devrait augmenter avec leur niveau d'expression et c'est bien ce qui est observé[125,139]. Ainsi chez des bactéries comme *Bacillus subtilis* , *Mycoplasma genitalium*, *Mycoplasma pneumoniae*, où la répartition des gènes est très fortement biaisée on peut voir une inversion du biais selon que l'on considère que les positions 3 des codons ou bien toutes les positions[125].

UNE PRESSION DE MUTATION EST ELLE SÉLECTIONABLE ?

Les effets d'une pression de mutation directionnelle ne se faisant sentir que par une lente modification des fréquences des informations génétiques dans une population, si sélection il y a ce n'est pas au niveau de **N** mais à une échelle de temps bien plus grande où ce sont plusieurs populations génétiques qui sont en compétition. Ce n'est pas impossible, la recombinaison chez les espèces diploïdes et mérodiploïdes et l'exemple même d'un processus que l'on pense avoir été sélectionné pour ses effets à long terme. Il y a-t-il des exemples de récupération adaptative des effets à long terme d'une pression de mutation directionnelle ?

Thermostabilité et isochores

Bernardi a proposé[16,13,14,15,17] que le taux élevé de base S observé dans certaines régions (les isochores lourds) des génomes des vertébrés à sang chaud puisse être sélectivement avantageux de par ses facultés thermostabilisantes soit directement au niveau de l'ADN, soit en augmentant l'hydrophobicité des protéines codées dans ces régions. Si cette hypothèse avait un quelconque caractère de généralité, on ne voit pas pourquoi on observe des isochores lourds chez *Crocodylus niloticus* et *Trachemys scripta elegans* qui sont des vertébrés à sang froid[72], ni pourquoi on observe une absence totale de corrélation entre le taux de bases S des génomes bactériens et leur température optimale de croissance[55], ni pourquoi chez les drosophiles les plus forts taux de bases S sont observés dans les espèces vivant dans des environnements froids[150]. Quant à l'interprétation de la corrélation entre taux de bases S et l'hydrophobicité des protéines[33], elle bénéficierait sans doute d'une meilleure structuration des données analysées en distinguant les protéines cytoplasmiques des protéines membranaires intégrales, premier facteur de la variabilité intraprotéomique[111]. J'ai représenté ci-dessous l'évolution de l'indice d'hydrophobicité moyen[100] des protéines chez 59 espèces bactériennes[111], comme on peut le constater si pression de sélection il y a, il faudrait expliquer pourquoi elle joue en sens inverse sur les protéines cytoplasmiques et les protéines membranaires intégrales.



Integral Membrane Proteins

```
y = ax + b; n=59; R^2 = 0.151
a = -0.155 [-0.252,-0.057]
b = +0.849 [+0.788,+0.909]
```

Cytoplasmic Proteins

```
y = ax + b; n=59; R^2 = 0.413
a = +0.303 [+0.207,+0.399]
b = -0.413 [-0.472,-0.354]
```

<S3>

L'hypothèse d'un avantage sélectif d'un fort taux en bases S lié à la température n'est donc pas un exemple convaincant de pression de mutation directionnelle ayant été sélectionnée.

Les codes génétiques

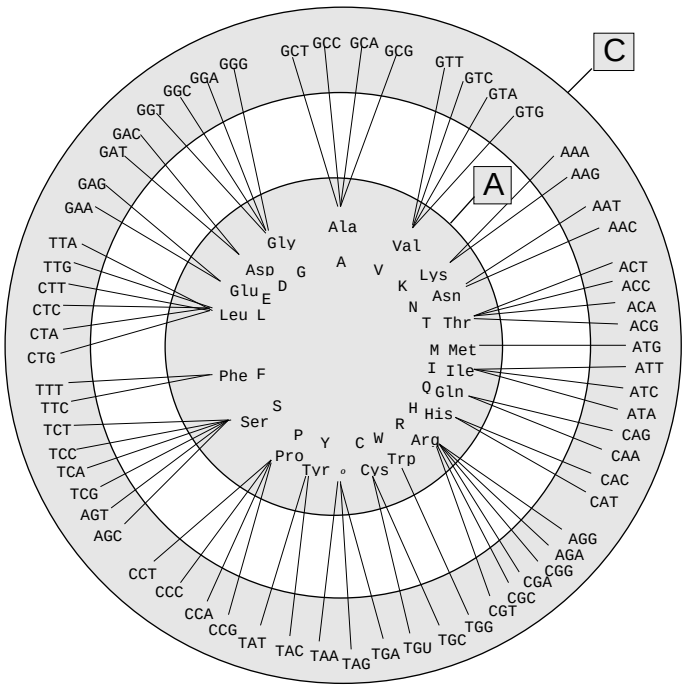
Soit **C** l'ensemble des 64 codons possibles dans les séquences codantes,

C = {AAA, AAC, AAG, AAT, ..., TTT },

et soit A l'ensemble formé de l'union de l'ensemble vide et des 20 acides aminés des protéines,

A = { , Ala, Arg, ..., Val },

où l'ensemble vide représente les codons non-sens utilisés comme fin de signal de la traduction et les codons non-assignés. Un code génétique est une application surjective de **C** dans **A** : tous les éléments de **C** ont une image dans A, et tous les éléments de A ont au moins un antécédent dans **C**, comme dans l'exemple ci-dessous correspondant au code génétique dit « universel ».



Les codes génétiques sont généralement représentés sous la forme d'un tableau 4x4 croisant les deux premières positions des codons et listant dans chaque case les quatre possibilités restantes pour la troisième position.

Le code génétique « universel »							
C	**A**	**C**	**A**	**C**	**A**	**C**	**A**
TTT	Phe	TCT	Ser	TAT	Tyr	TGT	Cys
TTC	Phe	TCC	Ser	TAC	Tyr	TGC	Cys
TTA	Leu	TCA	Ser	TAA		TGA	
TTG	Leu	TCG	Ser	TAG		TGG	Trp
CTT	Leu	CCT	Pro	CAT	His	CGT	Arg
CTC	Leu	CCC	Pro	CAC	His	CGC	Arg
CTA	Leu	CCA	Pro	CAA	Gln	CGA	Arg
CTG	Leu	CCG	Pro	CAG	Gln	CGG	Arg
ATT	Ile	ACT	Thr	AAT	Asn	AGT	Ser
ATC	Ile	ACC	Thr	AAC	Asn	AGC	Ser
ATA	Ile	ACA	Thr	AAA	Lys	AGA	Arg
ATG	Met	ACG	Thr	AAG	Lys	AGG	Arg
GTT	Val	GCT	Ala	GAT	Asp	GGT	Gly
GTC	Val	GCC	Ala	GAC	Asp	GGC	Gly
GTA	Val	GCA	Ala	GAA	Glu	GGA	Gly
GTG	Val	GCG	Ala	GAG	Glu	GGG	Gly

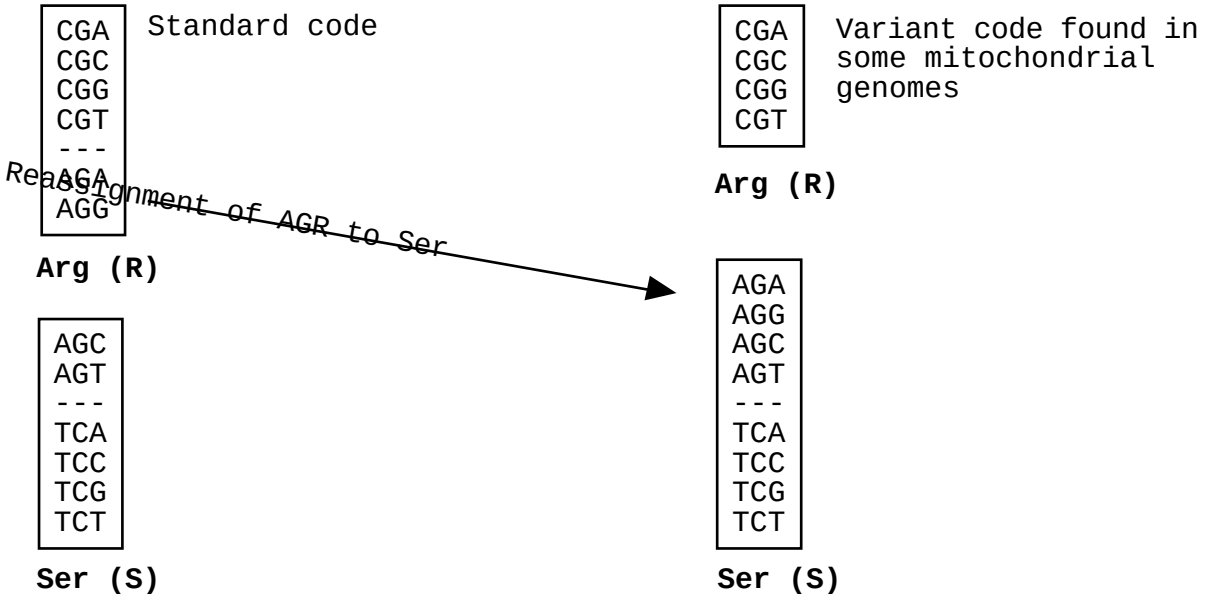
Une notation plus compacte des codes génétiques consiste à utiliser le code mono-lettre des acides aminés. Si l'on néglige les cas particuliers du codon d'initiation de la traduction, ainsi qu'un certain nombre d'exception telles que celle de la traduction de la sélénocystéine, un code génétique est donné par une chaîne de 64 caractères pouvant prendre 21 valeurs. Les codes génétiques connus sont représentés dans l'alignement ci-dessous, où les déviations par rapport au code génétique « universel » sont mises en évidence :

```
Base1  = TTTTTTTTTTTTTTTTCCCCCCCCCCCCCCCCAAAAAAAAAAAAAAAAGGGGGGGGGGGGGGGG
Base2  = TTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGGTTTTCCCCAAAAGGGG
Base3  = TCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAGTCAG

1        FFLLSSSSYY  CC WLLLLPPPPHHQQRRRRIIIMTTTTNNKKSSRRVVVVAAAADDEEGGGG
4        ..............W...............................................
10       ..............C...............................................
2        ..............W...................M...........  ..............
3        ..............W.TTTT..............M...............................
5        ..............W...................M...........SS..............
21       ..............W...................M.......N...SS..............
9        ..............W.........................N...SS..............
14       ..........Y...W.........................N...SS..............
13       ..............W...................M...........GG..............
16       ...........L...................................................
15       ...........Q...................................................
6        ..........QQ...................................................
12       ................S...............................................

constant XXXXXXXXXX..XX.X....XXXXXXXXXXXXXX.XXXXXXX.XXX..XXXXXXXXXXXXXXXX
```

1. The Standard Code. 4. The Mold, Protozoan, and Coelenterate Mitochondrial Code and the Mycoplasma/Spiroplasma Code 10. The Euplotid Nuclear Code **2**. The Vertebrate Mitochondrial Code **3**. The Yeast Mitochondrial Code **5**. The Invertebrate Mitochondrial Code **21**. Trematode Mitochondrial Code **9**. The Echinoderm Mitochondrial Code **14**. The Flatworm Mitochondrial Code **13**. The Ascidian Mitochondrial Code **16**. Chlorophycean Mitochondrial Code 15. Blepharisma Nuclear Code 6. The Ciliate, Dasycladacean and Hexamita Nuclear Code 12. The Alternative Yeast Nuclear Code

Les déviations par rapport au code universel ne concernent que 11 codons sur les 64. L'interprétation est que les codes variants dérivent du code universel par perte d'usage d'un codon sous l'influence d'une forte pression de mutation directionelle[132] ($\mu_D \to 0.0$ ou $\mu_D \to 1.0$), puis par désassignation et réassignation pour optimiser la traduction[2,4].

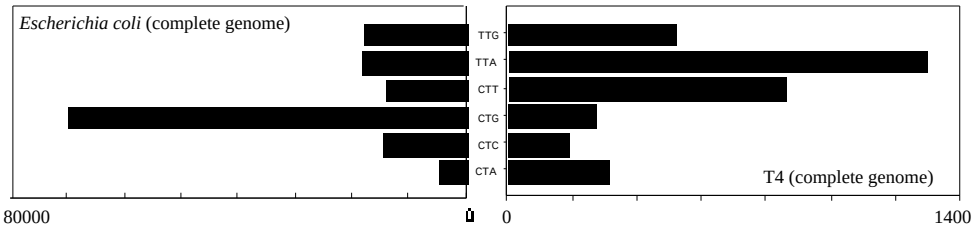
La mise en place d'un code génétique variant fait appel au mécanisme dit de capture de codon[78,132] : a) une pression de mutation directionnelle conduit à la perte d'usage d'un codon dans un génome b) une mutation dans l'ARNt reconnaissant ce codon le désassigne, mais comme il n'est pas reconnu par les facteurs de libération de la traduction, une mutation réverse vers ce codon est fortement contre sélectionnée pour ne pas bloquer la traduction. Cette situation intermédiaire existe chez *Micrococcus luteus* ($\hat{\mu}_D = 0.95$ [171]) où les codons AGA et ATA sont non assignés[84], chez *Mycoplasma capricolum* ($\hat{\mu}_D = 0.07$ [171]) où le codon CGG est non assigné[131], chez les mitochondries de *Balanoglossus carnosus* ($\hat{\mu}_D = 0.50$) où le codon AAA est non assigné[24] c) une mutation dans un ARNt permet de réassigner le codon. Par exemple, la très courante réassignation du codon stop TGA à Trp a été acquis indépendamment dans de nombreuses lignées[73], c'est un exemple de convergence évolutive au niveau moléculaire dû à une pression de mutation directionnelle. Andersson et Kurland ont suggéré[3,4] que la réassignation d'un codon pourrait avoir une valeur adaptative pour les génomes soumis à une forte pression sélective pour réduire leur taille comme chez les organelles où des bactéries intracellulaires où le processus de réduction du génome est en cours[1,2] comme en témoigne la très forte proportion de pseudogènes chez *Rickettsia prowazekii* (25 %) ou chez *Mycobacterium leprae*. À cause des 3 codons sextets dans le code génétique universel, le nombre minimum d'ARNt ne peut être inférieur à 23 dans un génome utilisant le code standard[132], il est troublant de constater que de nombreuses réassignations de codons permettent précisément de passer sous cette limite, comme la réassignation de AGR de Arg à Ser,



qui permet d'économiser un ARNt par rapport au code standard. Cependant, chez les mitochondries des Ascidiacea les codons AGR sont réassignés à Gly, ce qui ne permet pas d'économiser d'ARNt, mais probablement permet de lutter contre la perte d'usage des codons GGN standard pour Gly. De même, d'autres réassignations observées, comme de AAA de Lys vers Asn, ne permettent pas d'économiser d'ARNt. Ainsi, il n'est pas exclu que les effets d'une pression de mutation directionnelle aient été sélectivement avantageux dans quelques cas pour permettre des réassignations du code génétique permettant une meilleure efficacité de la traduction.

Le bactériophage T4

Les quasi-désassignations induits par le bactériophage T4 sont peut-être l'exemple d'une pression de mutation directionnelle dont les effets ont fini par être sélectivement avantageux. Le bactériophage T4 est soumis à une pression de mutation directionnelle ($\hat{\mu}_D = 0.22$ [89]) très différente de celle de son hôte *Escherichia coli* ($\hat{\mu}_D = 0.55$ [171]), ce qui n'est pas *a priori* sélectivement très avantageux puisqu'il oblige le bactériophage T4 à coder pour ses propres ARNt, il serait plus logique que le bactériophage T4 suive l'usage du code de son hôte pour optimiser sa traduction et compacter son génome. Par exemple, pour la Leucine nous avons les usages du code suivant pour *Escherichia coli* et son phage,



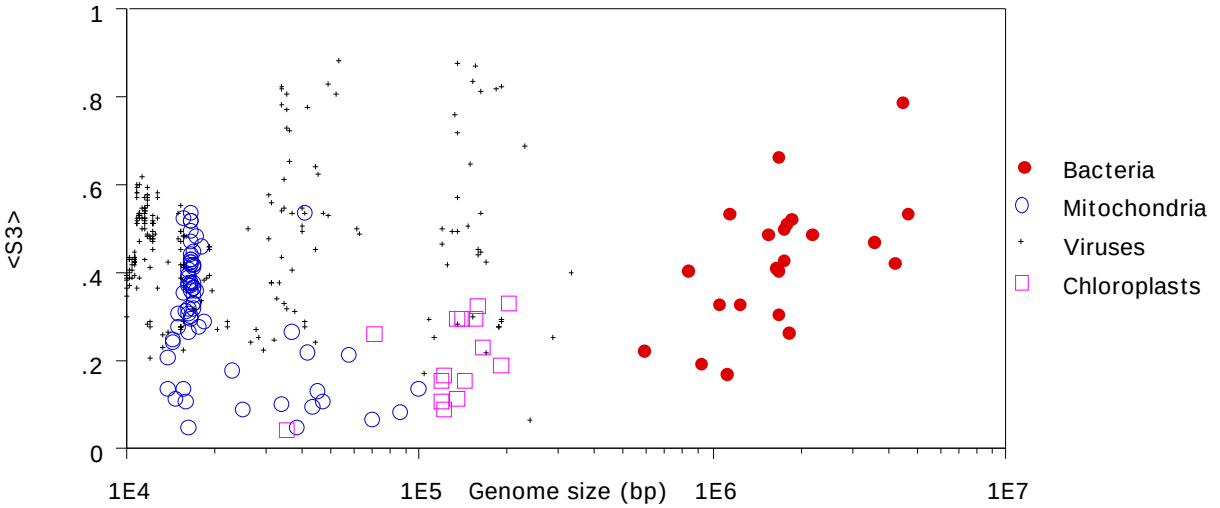
Ainsi, *Escherichia coli* utilise de façon préférentielle le codon CTG, c'est d'ailleurs le codon le plus fréquent, décodé par Leu-tRNA$_1$ qui présente une des plus grande concentration intracellulaire[74], alors que le bactériophage T4 préfère le codon TTA qui est décodé par son propre ARNt. On constate plus généralement que les codons décodables par les ARNt du phage sont fréquents dans les gènes du phage, et ce d'autant plus que les gènes sont exprimés tardivement dans le cycle d'infection du phage[31].Tamiko et Noboru Sueoka ont démontré[173,85,88,81,86,87,82] que moins de deux minutes après l'infection d'une cellule d'*Escherichia coli*, la concentration en Leu-tRNA$_1$ chute. C'est une quasi-désassignation induite par le phage dont le détail moléculaire est très complexe[95], mais il est clair qu'il est sélectivement avantageux pour le phage d'avoir un usage du code différent de son hôte. C'est le seul exemple à ma connaissance où l'on peut dire qu'une pression de mutation directionnelle est avantageuse a long terme.

Taux de mutation asymétrique

Furusawa et Doi ont montré[53,182,37,54] avec des simulations qu'un taux de mutation différent entre les deux brins de l'ADN est avantageux à long terme parce qu'il permet aux populations de supporter un fort taux de mutation tout en conservant une trace du génotype initial. Les pressions de mutation asymétriques seraient-elles un exemple d'un tel processus ? C'est une question ouverte, mais très spéculative.

Taux de bases S et taille des génomes

Les données sur les génomes complets ne permettent pas de mettre en évidence de relation particulière entre la taille des génomes et leur taux de bases S.



Il est vrai cependant que les organelles et les bactéries intracellulaires ont généralement de petits génomes et sont également pauvres en bases S[126,69], il n'y a par exemple pas de mitochondrie connue ayant plus de 55 % de bases S en position 3 des codons. Que la réduction de la taille de ces génomes s'accompagne de la perte d'enzyme de réparation de l'ADN, et donc à une plus forte sensibilité aux pressions de mutation directionnelles, semble logique. Mais on ne voit pas pourquoi les pressions de mutation seraient toujours dirigées dans le même sens.

Polarisation du chromosome de E. coli

Chez *Escherichia coli* le site *dif* localisé près du terminus de réplication est essentiel pour monomèriser les dimères chromosomiques provoqués par les recombinaisons homologues, environ 15 % des cellules en phase de croissance sont concernées, et les mutants *dif*$^-$ sont rapidement éliminés s'ils sont mis en compétition avec des individus *dif*$^+$, sauf s'ils sont également *recA*$^-$ et incapables de recombiner[136]. Ce qui est étonnant c'est la dépendance particulière du site *dif* à sa localisation sur le chromosome : son activité décroît progressivement lorsque l'on éloigne de sa région d'origine et n'est plus efficace à plus de 30kb, il n'est plus efficace si une inversion locale du chromosome est provoquée en amont ou en aval dans la zone d'activité de *dif*. On imagine donc volontiers des signaux asymétriques dans la zone d'activité de *dif* permettant de positionner correctement les sites *dif* au niveau du septum[30].

Ainsi, il y aurait une pression de sélection pour maintenir des signaux asymétriques dans la zone d'activité de *dif*. Mais des mutants qui ont perdu le site *dif* et la totalité de la zone d'activité de *dif* (jusqu'à au moins 155 kb en amont et 59 kb en aval) retrouvent un phénotype sauvage si l'on insère le site *dif* au point de jonction de la délétion ! L'activité de *dif* est donc indépendante de signaux spécifiques présents dans sa zone d'activité. Il est tentant de spéculer que *dif* réutilise des signaux asymétriques existants pour une autre raison. Salzberg et collaborateurs ont montré qu'un grand nombre d'oligomères asymétriques étaient distribués de façon biaisée entre les deux brins[155]. Faut-il y voir un effet de la pression de mutation asymétrique ? La composition globale du brin précoce chez *E. coli* (A=1137535, C=1140273, G=1215935, T=1145478) ne conduit qu'à un faible enrichissement en base K (50.9 %), le rapport du nombre de motifs K_n dans le brin précoce sur leur nombre dans le brin tardif, $(0.509/0.491)^n$, ne vaut que 1.33 pour des octamères. La pression de mutation asymétrique chez *Escherichia coli* semble donc trop faible pour avoir produit des oligomères asymétrique dont la répartition serait fortement biaisée entre les deux brins.

CONCLUSION ET PERSPECTIVES

> La biologie positive doit donc être envisagée comme ayant pour destination générale de rattacher constamment l'un à l'autre, dans chaque cas déterminé, le point de vue anatomique et le point de vue physiologique, ou, en d'autres termes, l'état statique et l'état dynamique. Cette relation perpétuelle constitue son vrai caractère philosophique.
>
> Auguste Compte
> Cours de philosophie positive
> 1840-1842

Si nous faisons l'hypothèse que l'évolution des fréquences des bases de l'ADN est gouvernée par un processus symétrique par rapport au deux brins de l'ADN, nous obtenons un joli petit modèle faux. Ce modèle est joli dans le sens où il peut être rejeté à partir de la seule inspection des fréquences des bases dans de l'ADN, et son rejet qui est effectif dans de nombreux génomes signifie le processus sous-jacent inobservable est asymétrique.

Dans certains cas favorables comme celui de *Borrelia burgdorferi* l'interprétation biologique est qu'il y a une pression de mutation asymétrique entre les deux brins de l'ADN, il semble en tout cas extrêmement difficile de trouver une interprétation sélectionniste convaincante de la structure en chirochore de son génome. L'universalité des biais provoqués par les pressions de mutation asymétrique suggère qu'un mécanisme assez général est à l'œuvre. La théorie de la désamination accélérée des cytosines dans de l'ADN monocaténaire est une candidate intéressante, mais elle est difficile à tester. Il me semble pourtant important de comprendre la raison de l'universalité des biais parce que ceux-ci sont très prononcés chez des bactéries pathogènes comme *Borrelia burgdorferi* (maladie de Lyme) *Chlamydia pneumoniae* (pneumonie) *Chlamydia trachomatis* (trachome) *Rickettsia prowazekii* (typhus) et *Treponema pallidum* (syphilis) et absents chez l'homme. Peut-on mettre au point des agents antimicrobiens spécifiquement ciblés contre les génomes présentant de forts biais ? Si c'est bien une mauvaise protection de l'ADN simple brin qui est à l'origine des biais, on peut alors imaginer que la transcription soit également ciblée, et donc l'ensemble du métabolisme. Mais les biais que nous observons sont le fruit d'une longue évolution, et c'est peut-être une toute petite différence qui en est à l'origine, une différence trop petite pour être exploitable sur une échelle de temps humaine. Quoi qu'il en soit, une première étape est de mieux comprendre la raison de l'universalité des biais ne serait-ce parce que les invariants sont si rares en biologie qu'il est insupportable ne pas en comprendre l'origine.

La modélisation (c'est-à-dire la traduction dans un système formel mathématique ou informatique) de la théorie de la désamination des cytosines ou de toute autre théorie explicative sera un problème majeur. En effet, d'après mon expérience, c'est une démarche extrêmement lente et pénible et coûteuse. L'interprétation des résultats sur les biais est loin d'être évidente, et sans une analyse théorique claire des résultats attendus sous un modèle donné je ne vois pas comment nous pourrions progresser dans la compréhension du phénomène. Nous sommes peut-être aveugles à des résultats expérimentaux déjà présents dans les banques. C'est pourquoi j'aimerais proposer des sujets de recherche ayant trait à la modélisation de l'évolution des fréquences des bases de l'ADN dans des conditions non symétriques.

REFERENCES

1. Andersson, J.O., Andersson, S.G.E. (1999) Insights into the evolutionary process of genome degradation. *Current Opinion in Genetics & Development*, **9**:664-671.

2. Andersson, J.O., Andersson, S.G.E. (1999) Genome degradation is an ongoing process in *Rickettsia*. *Molecular Biology and Evolution*, **16**:1178-1191.

3. Andersson, S.G.E., Kurland, C.G. (1990) Codon preferences in free-living microorganisms. *Microbiological Reviews*, **54**:198-210.

4. Andersson, S.G.E., Kurland, C.G. (1991) An extreme codon preference strategy: codon reassignment. *Molecular Biology and Evolution*, **8**:530-544.

5. Andersson, S.G.E., Zomorodipour, A., Andersson, J.O., Sicheritz-Pontérn, T., Alsmark, U.C.M., Podowski, R.M., Naslund, A.K., Eriksson, A.-S., Winkler, H.H., Kurland, C.G. (1998) The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature*, **396**:133-140.

6. Anonymous (1986) Nomenclature for incompletely specified bases in nucleic acid sequences. Recommendations 1984. Nomenclature Committee of the International Union of Biochemistry (NC-IUB). *Proceedings of the National Academy of Sciences of the United States of America*, **83**:4-8.

7. Anonymous (2000) Instructions to Authors. *Molecular Biology and Evolution*, **17**:207-212.

8. Asakawa, S., Kumazawa, Y., Araki, T., Himeno, H., Miura, K.-I., Watanabe, K. (1991) Strand-specific nucleotide composition bias in echinoderm and vertebrate mitochondrial genomes. *Journal of Molecular Evolution*, **32**:511-520.

9. Baker, T.A., Wickner, S.H. (1992) Genetics and enzymology of DNA replication in *Escherichia coli*. *Annual Review of Genetics*, **26**:447-477.

10. Barbour, A.G. (1993). Linear DNA of *Borrelia* species and antigenic variation. *Trends in Microbiology*, **1**:236-239.

11. Bell, S.J., Forsdyke, D.R. (1999) Accounting units in DNA. *Journal of Theoretical Biology*, **197**:51-61.

12. Berkhout, B., van Hemert, F.J. (1994) The unusual nucleotide content of the HIV RNA genome results in a biased amino acid composition of HIV proteins. *Nucleic Acids Research*, **22**:1705-1711.

13. Bernardi, G. (1989) The isochore organization of the human genome. *Annual Review of Genetics*, **23**:637-661.

14. Bernardi, G. (1993) The vertebrate genome: isochores and evolution. *Molecular Biology and Evolution*, **10**:186-204.

15. Bernardi, G. (2000) Isochores and the evolutionary genomics of vertebrates. *Gene*, **241**:3-17.

16. Bernardi, G. and Bernardi, G. (1986) Compositional constraints and genome evolution. *Journal of Molecular Evolution*, **24**:1-11.

17. Bernardi, G., Olofsson, B., Filipski, J., Zerial, M., Salinas, J., Cuny, G., Meunier-Rotival, M., Rodier, F. (1985) The mosaic genome of warm-blooded vertebrates. *Science*, **228**, 953-958.

18. Blattner, F.R., Plunkett III, G., Bloch, C.A., Perna, N.T., Burland, V., Rilley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B., Shao, Y. (1997) The complete genome sequence of *Escherichia coli* K-12. *Science*, **277**:1453-1462.

19. Brewer, B.J. (1988) When polymerase collide: replication and the transcriptional organization of the *E. coli* chromosome. *Cell*, **53**:679-686.

20. Bronson, E.C., Anderson, J.N. (1994) Nucleotide composition as a driving force in the evolution of retroviruses. *Journal of Molecular Evolution*, **38**:506-532.

21. Brown, G.G., Simpson, M.V. (1982) Novel features of animal mtDNA evolution as shown by sequences oftwo rat cytochrome oxidase subunit II genes. *Proceedings of the National Academy of Sciences of the United States of America*, **79**:3246-3250.

22. Bulmer, M. (1991). Strand symmetry of mutation rates in the ß-globin region. *Journal of Molecular Evolution*, **33**:305-310.

23. Casjens, S., Palmer, N., van Vugt, R., Huang, W.M., Stevenson, B., Rosa, P., Lathigra, R., Sutton, G., Peterson, J., Dodson, R.J., Haft, D., Hickey, E., Gwinn, M., White, O., Fraser, C.M. (2000) A bacteria genome in flux: the twelve linear and nine circular extrachromosomal DNAs in an infectious isolate of Lyme disease spirochete *Borrelia burgdorferi*. *Molecular Microbiology*, **35**:490-516.

24. Castresana, J., Feldmaier-Fuchs, G., Pääbo, S. (1998) Codon reassignment and amino acid composition in hemichordate mitochondria. *Proceedings of the National Academy of Sciences of the United States of America*, **95**:3703-3707.

25. Cebrat, S., Dudek, M.R. (1998) The effect of DNA phase structure on DNA walks. *The European Physical Journal B*, **3**:271-276.

26. Cebrat, S., Dudek, M.R., Gierlik, A., Kowalczuk, M., Mackiewicz, P. (1999) Effect replication on the third base of codons. *Physica A*, **265**:78-84.

27. Chargaff, E. (1979) How genetics got a chemical education. *Annals of the New York Academy of Sciences*, **325**:345-360.

28. Clark, M.A., Moran, N.A., Baumann, P. (1999) Sequence evolution in bacterial endosymbionts having extreme base compositions. *Molecular Biology and Evolution*, **16**:1586-1598.

29. Collins, D.W., Jukes, T.H. (1993) Relationship between G+C in silent sites of codons and amino acid composition of human proteins. *Journal of Molecular Evolution*, **36**:201-213.

30. Cornet, F., Louarn, J., Patte, J., Louarn, J.-M. (1996) Restriction of the activity of the recombination site *dif* to a small zone of the *Escherichia coli* chromosome. *Genes & Development*, **10**:1152-1161.

31. Cowe, E., Sharp, P.M. (1991) Molecular evolution of bacteriophages: discrete patterns of codon usage in T4 genes are related to the time of gene expression. *Journal of Molecular Evolution*, **33**:13-22.

32. Cruveiller, S., Jabbari, K., D'Onofrio, G., Bernardi, G. (1999) Different hydrophobicities of orthologous proteins from *Xenopus* and human. *Gene*, **238**:15-21.

33. D'Onofrio, G., Jabbari, K., Musto, H., Bernardi, G. (1999) The correlation of protein hydropathy with the base composition of coding sequences. *Gene*, **238**:3-14.

34. D'Onofrio, G., Mouchiroud, D., Aïssani, B., Gautier, C., Bernardi, G. (1991) Correlations between the compositional properties of human genes, codon usage, and amino acid composition of proteins. *Journal of Molecular Evolution*, **32**:504-510.

35. Danchin, A. (1998) The Delphic boat or what the genomic texts tell us. *Bioinformatics*, **14**:383-383.

36. Danchin, A. (1999) From function to sequence, an integrated view of the genome texts. *Physica A*, **273**:92-98.

37. Doi, H., Furusawa, M. (1996) Evolution is promoted by asymmetrical mutations in DNA replication - genetic algorithm with double-stranded DNA -. *FUJITSU Sci. Tech. J.*, **2**:248-255.

38. Fickett, J.W., Torney, D.C., Wolf, D.R. (1992) Base compositional structure of genomes. *Genomics*, **13**:1056-1064.

39. Filipski J (1990) Evolution of DNA sequences. Contributions of mutational bias and selection to the origin of chromosomal compartments. pp 1-54 in G. Obe (Ed.) Advances in mutagenesis research 2, Springer Verlag.

40. Forsdyke, D.R. (1995) Relative roles of primary sequence and (G+C)% in determining the hierarchy of frequencies of complementary trinucleotide pairs in DNAs of different species. *Journal of Molecular Evolution*, **41**:573-581.

41. Foster, D.M., Jacquez, J.A. (1975) Multiple zeros for eigenvalues and the multiplicity of traps of a linear compartmental system. *Mathematical Biosciences*, **26**:89-97.

42. Francino, M.P., Chao, L., Riley, M.A. and Ochman, H. (1996) Asymmetries generated by transcription-coupled repair in enterobacterial genes. *Science*, **272**:107-109.

43. Francino, M.P., Ochman, H. (1997) Strand asymetries in DNA evolution. *Trends in Genetics*, **13**:240-245.

44. Francino, M.P., Ochman, H. (1999) A comparative genomics approach to DNA asymmetry. *Annals of the New York Academy of Sciences*, **870**:428-431.

45. Francino, M.P., Ochman, H. (2000) Strand symmetry around the ß-globin origin of replication in primates. *Molecular Biology and Evolution*, **17**:416-422.

46. Frank, A.C., Lobry, J.R. (1999) Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene*, **238**:65-77.

47. Fraser, C.M., Casjens, S., Huang, W.M., Sutton, G.G., Clayton, R., Lathigra, R., White, O., Ketchum, K.A., Dodson, R., Hickey, E.K., Gwinn, M., Dougherty, B., Tomb, J.F., Fleischmann, R.D., Richardson, D., Peterson, J., Kerlavage, A.R., Quackenbush, J., Salzberg, S., Hanson, M., van Vugt, R., Palmer, N., Adams, M.D., Gocayne, J., Weidman, J., Utterback, T., Watthey, L., McDonald, L., Artiach, P., Bowman, C., Garland, S., Fuji, C., Cotton, M.D., Horst, K., Roberts, K., Hatch, B., Smith, H.O., Venter, J.C. (1997) Genomic sequence of a Lyme disease spirochaete, *Borrelia burgdorferi*. *Nature*, **390**:580-586.

48. Fraser, C.M., Norris, S.J., Weinstock, G.M., White, O., Sutton, G.G., Dodson, R., Gwinn, M., Hickey, E.K., Clayton, R., Ketchum, K.A., Sodergren, E., Hardham, J.M., McLeod, M.P., Salzberg, S., Peterson, J., Khalak, H., Richardson, D., Howell, J.K., Chidambaram, M., Utterback, T., McDonnald, L., Artiach, P., Bowman, C., Cotton, M.D., Fujii, C., Garland, S., Hatch, B., Horst, K., Roberts, K., Sandusky, M., Weidman, J., Smith, H.O., Venter, J.C. (1998) Complete genome sequence of

Treponema pallidum, the syphilis spirochete. *Science*, **281**:375-388.

49. Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M., Fritchman, J.L., Weidman, J.F., Small, K.V., Sandusky, M., Fuhrmann, J., Nguyen, D., Utterback, T.R., Saudek, D.M., Phillips, C.A., Merrick, J.M., Tomb, J.-F., Dougherty, B.A., Bott, K.F., Hu, P.-C., Lucier, T.S., Peterson, S.N., Smith, H.O., Hutchison, C.A., Venter, J.C. (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**:397-403.

50. Frederico, L.A., Kunkel, T.A., Shaw, B.R. (1990) A sensitive genetic assay for the detection of cytosine deamination: determination of rate constants and the activation energy. *Biochemistry*, **29**:2532-2537.

51. Freeman, J.M., Plasterer, T.N., Smith, T.F., Mohr, S.C. (1998) Patterns of genome organization in bacteria. *Science*, **279**:1827.

52. Freese, E. (1962) On the evolution of the base composition of DNA. *Journal of Theoretical Biology*, **3**:82-101.

53. Furusawa, M., Doi, H. (1992) Promotion of evolution: disparity in the frequency of strand-specific misreading between the lagging and leading DNA strands enhances disproportionate accumulation of mutations. *Journal of Theoretical Biology*, **157**:127-133.

54. Furusawa, M., Doi, H. (1998) Asymmetrical DNA replication promotes evolution: disparity theory of evolution. *Genetica*, **102/103**:333-347.

55. Galtier, N., Lobry, J.R. (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of Molecular Evolution*, **44**:632-636.

56. Gillespie, J.H. (1991) The causes of molecular evolution. Oxford University press. ISBN 0-19-506883-1.

57. Gouy, M., Gautier, C. (1982) Codon usage in bacteria: correlation with gene expressivity. *Nucleic Acids Research*, **10**:7055-7073.

58. Gouy, M., Milleret F, Mugnier C, Jacobzone M, Gautier C (1984) ACNUC: a nucleic acid sequence data base and analysis system. *Nucleic Acids Research*, **12**:121-127.

59. Grantham, R., Gautier, C. (1980) Genetic distances from mRNA sequences. *Naturwissenschaften*, **67**:93-94.

60. Grantham, R., Gautier, C., Gouy, M. (1980) Codon frequencies in 119 individual genes confirm consistent choices of degenerate bases according to genome type. *Nucleic Acids Research*, **8**:1893-1912.

61. Grantham, R., Gautier, C., Gouy, M., Mercier, R. and Pavé, A. (1980) Codon catalog usage and the genome hypothesis *Nucleic Acids Research*, **8**:r49-r62.

62. Graur, D., Li, W.-H. (2000) Fundamentals of molecular evolution, Second Edition. Sinauer Associates Inc., Sunderland, Massachusetts, USA. ISBN 0-87893-266-6.

63. Grigoriev, A. (1998) Analyzing genomes with cumulative skew diagrams. *Nucleic Acids Research*, **26** :2286-2290.

64. Grigoriev, A. (1999) Strand-specific compositional asymmetries in double-stranded DNA viruses. *Virus Research*, **60**:1-19.

65. Grigoriev, A., Freeman, J.M., Plasterer, T.N., Smith, T.F., Mohr, S.C. (1998) Genome arithmetic. *Science* **281**:1923-1924.

66. Gu, X., Hewett-Emmett, D., Li, W.-H. (1998) Directional mutational pressure affects the amino acid composition and hydrophobicity of proteins in bacteria. *Genetica*, **102/103**:383-391.

67. Hanai, R., Wada, A. (1988) The effects of guanine and cytosine variation on dinucleotide frequency and amino acid composition in the human genome. *Journal of Molecular Evolution*, **27**:321-325.

68. Haney, P.J., Badger, J.H., Buldak, G.L., Teich, C.I., Woese, C.R., Olsen, G.J. (1999) Thermal adaptation analyzed by comparison of protein sequences from mesophilic and extremely thermophilic *Methanococcus* species. *Proceedings of the National Academy of Sciences of the United States of America*, **96**:3578-3583.

69. Heddi, A., Charles, H., Khatchadourian, C., Bonnot, G., Nardon, P. (1998) Molecular characterization of the principal symbiotic bacteria of the weevil *Sitophilus oryzae*: a peculiar G+C content of an endocytobiotic DNA. *Journal of Molecular Evolution*, **47**:52-61.

70. Holm, L. (1986) Codon usage and gene expression. *Nucleic Acids Research*, **27**:244-247.

71. Holmes,W.M., Goldman,E., Miner,T.A. and Hatfield,G.W. (1977) *Proceedings of the National Academy of Sciences of the United States of America*, **74**:1393-1397.

72. Hughes, S., Zelus, D., Mouchiroud, D. (1999) Warm-blooded isochore structure in nile crocodile and turtle. *Molecular Biology and Evolution*, **16**:1521-1527.

73. Inagaki, Y., Ehara, M., Watanabe, K.I., Hayashi-Ishimaru, Y., Ohama, T. (1998) Directionally evolving genetic code: the

UGA codon from stop to tryptophan in mitochondria. *Journal of Molecular Evolution*, **47**:378-384.

74. Ikemura,T. (1981) Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes. *Journal of Molecular Biology*, **146**:1-21.

75. Jacquez, J.A., Simon, C.P. (1993) Qualitative theory of compartmental systems. *SIAM Review*, **35**:43-79

76. Jermiin, L.S., Graur, D., Crozier, R.H. (1995) Evidence from analyses of intergenic region for strand-specific directional mutation pressure in metazoan mitochondrial DNA. *Molecular Biology and Evolution*, **12**:558-563.

77. Jermiin, L.S., Graur, D., Lowe, R.M., Crozier, R.H. (1994) Analysis of directional mutation pressure and nucleotide content in mitochondrial cytochrome *b* genes. *Journal of Molecular Evolution*, **39**:160-173.

78. Jukes, T.H. (1985) A change in the genetic code in *Mycoplasma capricolum*. *Journal of Molecular Evolution*, **22**:361-362.

79. Jukes, T.H., Bhushan, V. (1986) Silent nucleotide substitutions and G+C content of some mitochondrial and bacterial genes. *Journal of Molecular Evolution*, **24**:39-44.

80. Jukes, T.H., Holmquist, R. and Moise, H. (1975) Amino acid composition of proteins: selection against the genetic code. *Science* **189**:50-51.

81. Kan, J., Kano-Sueoka, T., Sueoka, N. (1968) Characterization of leucine transfer ribonucleic acid in *Escherichia coli* following infection with bacteriophage T2. *Journal of Biological Chemistry*, **243**:5584-5590.

82. Kan, J., Nirenberg, M., Sueoka, N. (1970) Coding specificity of *Escherichia coli* leucine tRNA before and after infection with bacteriophage T2. *Journal of Molecular Biololgy*, **52**:179-193.

83. Kanaya, S., Yamada, Y., Kudo, Y., Ikemura, T. (1999) Studies of codon usage and tRNA genes of 18 unicellular organisms and quantification of *Bacillus subtilis* tRNA: gene expression level and species-specific diversity of codon usage based on multivariate analysis. *Gene*, **238**:143-155.

84. Kano, A., Ohama, T., Abe, R., Osawa, S. (1993) Unassigned or nonsense codons in *Micrococcus luteus*. *Journal of Molecular Biology*, **230**:51-56.

85. Kano-Sueoka, T., Sueoka, N. (1966) Modifications of leucyl-sRNA after bacteriophage infection. *Journal of Molecular Biology*, **20**:183-209.

86. Kano-Sueoka, T., Sueoka, N. (1968) Characterization of a modified leucyl-tRNA of *Escherichia coli* after bacteriophage T2 infection. *Journal of Molecular Biology*, **37**:475-491.

87. Kano-Sueoka, T., Sueoka, N. (1969) Leucine tRNA and cessation of *Escherichia coli* protein synthesis upon phage T2 infection. *Proceedings of the National Academy of Sciences of the United States of America*, **62**:1229-1236.

88. Kano-Sueoka, T., Nirenberg, M., Sueoka, N. (1968) Effect of bacteriophage infection upon the specificity of leucine transfer RNA for RNA codewords. *Journal of Molecular Biology*, **35**:1-12.

89. Kano-Sueoka, T., Lobry, J.R., Sueoka, N. (1999) Intra-strand biases in bacteriophage T4 genome. *Gene*, **238**: 59-64.

90. Karkas, J.D., Rudner, R., Chargaff, E. (1968) Separation of *B. subtilis* DNA into complementary strands, II. Template functions and composition as determined by transcription with RNA polymerase. *Proceedings of the National Academy of Sciences of the United States of America*, **60**:915-920.

91. Karkas, J.D., Rudner, R., Chargaff, E. (1970) Template properties of complementary fractions of denatured microbial deoxyribonucleic acids. *Proceedings of the National Academy of Sciences of the United States of America*, **65**:1049-1056.

92. Karlin, S. (1999) Bacterial DNA strand compositional asymmetry. *Trends in Microbiology*, **7**:305-308.

93. Karlin, S., Blaisdell, B.E., Schachtel, G.A. (1990) Contrasts in codon usage of latent versus productive genes of Epstein-Barr virus: data and hypotheses. *Journal of Virology*, **64**:4264-4273.

94. Karlin, S., Campbell, A.M., Mrázek, J. (1998) Comparative DNA analysis across diverse genomes. *Annual Review of Genetics*, **23**:185-225.

95. Kaufmann, G. (2000) Anticodon nucleases. *Trends in Biochemical Sciences*, **25**:70-74.

96. Kimura, M. (1968) Evolutionary rate at the molecular level. *Nature*, **217**:624-626.

97. Kimura, M. (1985) Diffusion models in population genetics with special reference to fixation time of molecular mutants under mutational pressure. In Population Genetics and Molecular Evolution, pp. 19-39 ed. T. Ohta and K. Aoki. Tokyo: Japan Scientific Societies Press / Berlin: Springer-Verlag.

98. King, J.L., Jukes, T.H. (1969) Non-Darwinian evolution. *Science*, **164**:788-798.

99. Kunst, F., Ogasawara, N., Moszer, I., Albertini, A.M., Alloni, G., Azevedo, V., Bertero, M.G., Bessières, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignell, S.C., Bron, S., Brouillet, S., Bruschi, C.V., Caldwell, B., Capuano, V., Carter, N.M., Choi, S.-K., Codani, J.-J., Connerton, I.F., Cummings, N.J., Daniel, R.A., Denizot, F., Devine, K.M., Düsterhöft, A., Ehrlich, S.D., Emmerson, P.T., Entian, K.D., Errington, J., Fabret, C., Ferrari, E., Foulger, D., Fritz, C., Fujita, M., Fujita, Y., Fuma, S., Galizzi, A., Galleron, N., Ghim, S.-Y., Glaser, P., Goffeau, A., Golightly, E.J., Grandi, G., Guiseppi, G., Guy, B.J., Haga, K., Haiech, J., Harwood, C.R., Hénaut, A., Hilbet, H., Holsappel, S., Hosono, S., Hullo, M.-F., Itaya, M., Jones, L., Joris, B., Karamata, D., Kasahara, Y., Klaerr-Blanchard, M., Klein, C., Kobayashi, Y., Koetter, P., Koningstein, G., Krogh, S., Kumano, M., Kurita, K., Lapidus, A., Lardinois, S., Lauber, J., Lazarevic, V., Lee, S.-M., Levine, A., Liu, H., Masuda, S., Mauël, C., Médigue, C., Medina, N., Mellado, R.P., Mizuno, M., Moestl, D., Nakai, S., Noback, M., Noone, D., O'Reilly, M., Ogawa, K., Ogiwara, A., Oudega, B., Park, S.-H., Parro, V., Pohl, T.M., Portetelle, D., Porwollik, S., Prescott, A.M., Presecan, E., Pujic, P., Purnelle, B., Rapoport, G., Rey, M., Reynolds, S., Rieger, M., Rivolta, C., Rocha, E., Roche, B., Rose, M., Sadaie, Y., Sato, T., Scanlan, E., Schleich, S., Schroeter, R., Scoffone, F., Sekiguchi, J., Sekowska, A., Serror, S.J., Serror, P., Shin, B.-S., Soldo, B., Sorokin, A., Tacconi, E., Takagi, T., Takahashi, H., Takemaru, K., Takeuchi, M., Tamakoshi, A., Tanakata, T., Terpstra, P., Tognoni, A., Tosato, V., Uchiyama, S., Vandenbol, M., Vannier, F., Vassarotti, A., Viari, A., Wambutt, R., Wedler, E., Wedler, H., Weitzenegger, T., Winters, P., Wipat, A., Yamamoto, H., Yamane, K., Yasumoto, K., Yata, K., Yoshida, K., Yoshikawa, H.-F., Zumstein, E., Yoshikawa, H., Danchin, A. (1997) The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*. *Nature*, **390**:249-256.

100. Kyte, J., Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, **157**:105-132.

101. Lafay, B., Lloyd, A.T., McLean, M.J., Devine, K.M., Sharp, P.M., Wolfe, K.H. (1999) Proteome composition and codon usage in spirochaetes: species-specific and DNA strand-specific mutational biases. *Nucleic Acids Research*, **27**:1642-1649.

102. Li, W. (1999) Statistical properties of open reading frames in complete genome sequences. *Computer & Chemistry* **23**:283-301.

103. Li, W.-H. (1987) Models of nearly neutral mutations with particular implications for nonrandom usage of synonymous codons. *Journal of Molecular Evolution*, **24**:337-345.

104. Li, W.-H. (1997). Molecular evolution. Sinauer Associates, Sunderland Massachusetts U.S.A.

105. Liu, B., Alberts, M. (1995) Head-on collision between a DNA replication apparatus and RNA polymerase transcription complex. *Science*, **267**:1131-1137.

106. Liò, P., Goldman, N. (1998) Models of molecular evolution and phylogeny. *Genome Research*, **8**:1233-1244.

107. Lobry, J.R. (1995) Properties of a general model of DNA evolution under no-strand bias conditions. *Journal of Molecular Evolution*, **40**:326-330 ; **41**:680.

108. Lobry, J.R. (1996) A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biochimie*, **78**:323-326.

109. Lobry, J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular Biology and Evolution*, **13**:660-665.

110. Lobry, J.R. (1996) Origin of replication of *Mycoplasma genitalium. Science*, **272**:745-746.

111. Lobry, J.R. (1997) Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. *Gene*, **205**:309-316.

112. Lobry, J.R. (1999) A nice wrong model for the evolution of DNA base frequencies. *Physica A*, **273**:99-102.

113. Lobry, J.R. (1999) Genomic landscapes. *Microbiology Today*, **26**:164-165.

114. Lobry, J.R., Sueoka, N. (2000) Asymmetric directional mutation pressures in bacteria. *in preparation*.

115. Lobry, J. R., Gautier, C. (1994) Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Research*, **22**:3174-3180.

116. Lobry, J.R., Lobry, C. (1999) Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. *Molecular Biology and Evolution*, **16**:719-723.

117. Lopez, P., Philippe, H., Myllykallio, H., Forterre, P. (1999) Identification of putative chromosomal origins of replication in Archaea. *Molecular Microbiology*, **32**:883-891.

118. Lyamichev, V., Panyutin, I., Frank-Kamenetskii, M.D. (1984) The absence of cruciform structure from pA03 plasmid

DNA *in vivo*. *Journal of Biomolecular Structure and Dynamics*, **2**:291-301.

119. Mackiewicz, P., Gierlik, A., Kowalczuk, M., Dudek, M.R., Cebrat, S. (1999) How does replication-associated mutational pressure influence amino acid composition of proteins? *Genome Research*, **9**:409-416.

120. Mackiewicz, P., Gierlik, A., Kowalczuk, M., Dudek, M.R., Cebrat, S. (1999) Asymmetry of nucleotide composition of prokaryotic chromosomes. *Journal of Applied Genetics*., **40**:1-14.

121. Mackiewicz, P., Gierlik, A., Kowalczuk, M., Szczepanik, D., Dudek, M.R., Cebrat, S. (1999) Mechanisms generating long-range correlation in nucleotide composition of the *Borrelia burgdorferi* genome. *Physica A*, **273**:103-115.

122. Marians, K.J. (1992) Prokaryotic DNA replication. *Annual Review of Biochemistry*, **61**:673-719.

123. McDonald, J.H., Grasso, A.M., Rejto, L.K. (1999) Patterns of temperature adaptation in proteins from *Methanococcus* and *Bacillus*. *Molecular Biology and Evolution*, **16**:1785-1790.

124. McInerney, J.O. (1998) Replication and transcriptional selection on codon usage in *Borrelia burgdorferi*. *Proceedings of the National Academy of Sciences of the United States of America*, **95**:10698-10703.

125. McLean, M.J., Wolfe, K.H., Devine, K.M. (1998) Base composition skews, replication orientation, and gene orientation in 12 prokaryote genomes. *Journal of Molecular Evolution* **47**:691-696.

126. Moran, N. (1996) Accelerated evolution and muller's rachet in endosymbiotic bacteria. *Proceedings of the National Academy of Sciences of the United States of America*, **93**:2873-2878.

127. Morton, B.R. (1999) Strand asymmetry and codon usage bias in the chloroplast genome of *Euglena gracilis*. *Proceedings of the National Academy of Sciences of the United States of America*, **96**:5123-5128

128. Mrázek, J., Karlin, S. (1998) Strand compositional asymmetry in bacterial and large viral genomes. *Proceedings of the National Academy of Sciences of the United States of America*, **95**:3720-3725.

129. Nossal, N.G. (1983) Prokaryotic DNA replication systems. *Annual Review of Biochemistry*, **52**:581-615.

130. Nussinov, R. (1982) Some indications for inverse DNA duplication. *Journal of Theoretical Biology*, **95**:783-791.

131. Oba, T., Andachi, Y., Muto, A., Osawara, S. (1991) CGG: an unassigned codon in *Mycoplasma capricolum*. *Proceedings of the National Academy of Sciences of the United States of America*, **88**:921-925.

132. Osawa, S., Jukes, T.H., Watanabe, K., Muto, A. (1992) Recent evidence for evolution of the genetic code. *Microbiological Review*s, **56**:229-264.

133. Panyutin, I., Jlishko, V., Lyamichev, V. (1984) Kinetics of cruciform formation and stability of cruciform structure in superhelical DNA. *Journal of Biomolecular Structure and Dynamics*, **1**:1311-1324.

134. Parkhill, J., Achtman, M., James, K.D., Bentley, S.D., Churcher, C., Klee, S.R., Morelli, G., Basham, D., Brown, D., Chillingworth, T., Davies, R.M., Davis, P., Devlin, K., Feltwell, T., Hamlin, N., Holroyd, S., Jagels, K., Leather, S., Moule, S., Mungall, K., Quail, M.A., Rajandream, M.-A., Rutherford, K.M., Simmonds, M., Skelton, J., Whitehead, S., Spratt, B.G., Barrell, B.G. (2000) Complete DNA sequence of a serogroup A strain of *Neisseria meningitidis* Z2491. *Nature*, **404**:502-506.

135. Parkhill, J., Wren, B.W., Mungall, K., Ketley, J.M., Churcher, C., Basham, D., Chillingworth, T., Davies, R.M., Feltwell, T., Holroyd, S., Jagels, K., Karlyshev, A.V., Moule, S., Pallen, M.J., Penn, C.W., Quail, M.A., Rajandream, M.-A., Rutherford, K.M., van Vhet, A.H.M., Withehead, S., Barrell, B.G. (2000) The genome sequence of the food-borne pathogen *Campylobacter jejuni* reveals hypervariable sequences. *Nature*, **403**:665-668.

136. Pérals, K., Cornet, F., Merlet, Y., Delon, I., Louarn, J.-M. (2000) Functional polarization of the *Escherichia coli* chromosome terminus: the *dif* site acts in chromosome dimer resolution only when located between long stetches of opposite polarity. *Molecular Microbiology*, **36**:33-43.

137. Perna, N.T., Kocher, T.D. (1995) Patterns of nucleotide composition at fourfold degenerate site of animal mitochondria genomes. *Journal of Molecular Evolution*, **41**:353-358.

138. Perrière, G., Bessières, P., Labedan, B. (2000) EMGLib: the enhanced microbial genomes library (update 2000). *Nucleic Acids Research*, **28**:68-71.

139. Perrière, G., Lobry, J.R. (1998) Asymmetrical coding sequence repartition and codon adaptation index values between leading and lagging strands in seven bacterial species. *in* Proceedings of the first international conference on bioinformatics of genome regulation and structure (Novosibirsk, Russia, August 24-31, 1998) **2**:254-255.

140. Perrière, G., Lobry, J.R., Thioulouse, J. (1996) Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acids sequences. *Computer Applications in the Biosciences*, **12**:519-524.

141. Picardeau, M., Lobry, J.R., Hinnebusch, B.J. (1999) Physical mapping of an origin of bidirectional replication at the center of the *Borrelia burgdorferi* linear chromosome. *Molecular Microbiology*, **32**:437-445.

142. Prabhu, V.V. (1993) Symetry observation in long nucleotide sequences. *Nucleic Acids Research*, **21**:2797-2800.

143. Rand, D.M., Kann, L.M. (1998) Mutation and selection at silent and replacement sites in the evolution of animal mitochondrial DNA. *Genetica* **102/103**:393-407.

144. Read, T.D., Brunham, R.C., Shen, C., Gill, S.R., Heidelberg, J.F., White, O., Hickey, E.K., Peterson, J., Utterback, T., Berry, K., Bass, S., Linher, K., Weidman, J., Khouri, H., Craven, B., Bowman, C., Dodson, R., Gwinn, M., Nelson, W., DeBoy, R., Kolonay, J., McClarty, G., Salzberg, S.L., Eisen, J., Fraser, C.M. (2000) Genome sequences of *Chlamydia trachomatis* MoPn and *Chlamydia pneumoniae* AR39. *Nucleic Acids Research*, **28**:1397-1406.

145. Reyes, A., Gissi, C., Pesole, G., Saccone, C. (1998) Asymmetrical directional mutation pressure in the mitochondrial genome of mammals. *Molecular Biology and Evolution*, **15**:957-966.

146. Rocha, E.P.C., Danchin, A., Viari, A. (1999) Bacterial DNA strand compositional asymmetry: Response. *Trends in Microbiology*, **7**:308-308.

147. Rocha, E.P.C., Danchin, A., Viari, A. (1999) Universal replication biases in bacteria. *Molecular Microbiology*, **32**:11-16.

148. Rocha, E.P.C., Viari, A., Danchin, A. (1998) Oligonucleotide bias in *Bacillus subtilis* : general trends and taxonomic comparisons. *Nucleic Acids Research*, **26**:2971-2980.

149. Rodríguez F., Olivier, J.L., Marín, A., Medina, J.R. (1990) The general stochastic model of nucleotide substitution. *Journal of Theoretical Biology*, **142**:485-501.

150. Rodríguez -Trelles, F., Tarrío, R., Ayala, F.J. (2000) Fluctuating mutation bias and the evolution of base composition in *Drosophila*. *Journal of Molecular Evolution*, **50**:1-10.

151. Rosso, L., Lobry, J.R., Flandrois, J.P. (1993) An unexpected correlation between cardinal temperatures of microbial growth highlighted by a new model. *Journal of Theoretical Biology*, **162**:447-463.

152. Rudner, R., Karkas, J.D., Chargaff, E. (1968) Separation of *B. subtilis* DNA into complementary strands, III. Direct analysis. *Proceedings of the National Academy of Sciences of the United States of America*, **60**:921-922.

153. Rudner, R., Karkas, J.D., Chargaff, E. (1969) Separation of microbial deoxyribonucleic acids into complementary strands. *Proceedings of the National Academy of Sciences of the United States of America*, **63**:152-159.

154. Rudner, R., LeDoux, M. (1974) Distribution of pyrimidine oligonucleotides in complementary strand fractions of *Escherichia coli* desoxyribonucleic acid. *Biochemistry*, **13**:118-125.

155. Salzberg, S.L., Salzberg, A.J., Kerlavage, A.R., Tomb, J.-F. (1998) Skewed oligomers and origins of replication. *Gene*, **217**:57-67.

156. Sharp, P.M., Matassi, G. (1994) Codon usage and genome evolution. *Current Opinion in Genetics and Development*, **4**:851-860.

157. Sharp, P.M., Li, W.-H. (1987) The codon adaptation index - a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Research*, **15**:1281-1295.

158. Sharp, P.M., Tuohy, T.M.F., Mosurski, K.R. (1986) Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes. *Nucleic Acids Research*, **14**:5125-5143.

159. Shpaer, E.G. (1989) Amino acid composition is correlated with protein abundance in Escherichia coli: can this be due to optimization of translational efficiency? *Protein Sequences and Data Analysis*, **2**:107-110.

160. Sinden, R.R., Broyles, S.S., Pettijohn, E. (1983) Perfect palindromic *lac* operator DNA sequence exists as a stable cruciform strucutre in supercoiled DNA *in vitro* but not *in vivo*. *Proceedings of the National Academy of Sciences of the United States of America*, **80**:1797-1801.

161. Sueoka, N. (1959) A statistical analysis of deoxyribonucleic acid distribution in density gradient centrifugation. *Proceedings of the National Academy of Sciences of the United States of America*, **45**:1480-1490.

162. Sueoka, N. (1961) Correlation between base composition of deoxyribonucleic acid and amino acid composition of protein. *Proceedings of the National Academy of Sciences of the United States of America*, **47**:1141-1149.

163. Sueoka, N. (1961) Variation and heterogeneity of base composition of deoxyribonucleic acids: a compilation of old

and new data. *Journal of Molecular Biology*, **3**:31-40.

164. Sueoka, N. (1962) On the genetic basis of variation and heterogeneity of DNA base composition. *Proceedings of the National Academy of Sciences of the United States of America*, **48**:582-592.

165. Sueoka, N. (1964) On the evolution of informational macromolecules. pp 479-496 in Bryson, V. and Vogel, H.J. (eds), *Evolving genes and proteins*. Academic Press, New York, USA.

166. Sueoka, N. (1988) Directional mutation pressure and neutral molecular evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **85**:2653-2657.

167. Sueoka, N. (1992) Directional mutation pressure, selective constraints, and genetic equilibria.*Journal of Molecular Evolution*, **34**:95-114.

168. Sueoka, N. (1993) Directional mutation pressure, mutator mutations, and dynamics of molecular evolution. *Journal of Molecular Evolution*, **37**:137-153.

169. Sueoka, N. (1995) Intrastrand parity rules of DNA base composition and usages biases of synonymous codons. *Journal of Molecular Evolution*, **40**:318-325; **42**:323.

170. Sueoka, N. (1999) Translation-coupled violation of parity rule 2 in human genes is not the cause of heterogeneity of the DNA G+C content of third codon position. *Gene*, **238**:53-58.

171. Sueoka, N. (1999) Two aspects of DNA base composition: G+C content and translation-coupled deviation from intra-strand rule of A = T and G = C. *Journal of Molecular Evolution*, **49**:49-62.

172. Sueoka, N., Marmur, J. and Doty, P. (1959) Heterogeneity in deoxyribonucleic acids. II. Dependence of the density of deoxyribonucleic acids on guanine-cytosine. *Nature* , **183**:1427-1431.

173. Sueoka, N., Kano-Sueoka, T. (1964) A specific modification of leucyl-sRNA of *Escherichia coli* after phage T2 infection. *Proceedings of the National Academy of Sciences of the United States of America*, **52**:535-1540.

174. Szybalski, W., Kubinski, H., Sheldrick, P. (1966) Pyrimidine clusters on the transcribing strand of DNA and their possible role in the initiation of RNA synthesis. *Cold Sping Harbor Symposium on Quantitative Biology*, **31**:123-127.

175. Tanaka, M., Ozawa, T. (1994) Strand asymmetry in human mitochondrial DNA mutations. *Genomics*, **22**:327-335.

176. Tettelin, H., Saunders, N.J., Heidelberg, J., Jeffries, A.C., Nelson, K.E., Eisen, J.A., Ketchum, K.A., Hood, D.W., Peden, J.F., Dodson, R.J., Nelson, W.C., Gwinn, M.L., DeBoy, R., Peterson, J.D., Hickey, E.K., Haft, D.H., Salzberg, S.L., White, O., Fleischmann, R.D., Dougherty, B.A., Mason, T., Ciecko, A., Parksey, D.S., Blair, E., Cittone, H., Clark, E.B., Cotton, M.D., Utterback, T.R., Khouri, H., Qin, H., Vamathevan, J., Gill, J., Scarlato, V., Masignani, V., Pizza, M., Grandi, G., Smith, H.O., Fraser, C.M., Moxon, E.R., Rappuoli, R., Venter, J.C. (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science*, **287**:1809-1815.

177. Thioulouse, J., Dolédec, S., Chessel, D. and Olivier, J.M. (1995) ADE software multivariate analysis and graphical display of environmental data. In Guariso,G. and Rizzoli,A. (eds), *Software per l'Ambiente*. Pàtron, Bolonia, pp. 57-62.

178. Thioulouse, J., Lobry, J.R. (1995) Co-inertia alaysis of amino-acid physico-chemical properties and protein composition with the ADE package. *Computer Applications in the Biosciences*, **11**:321-329.

179. Tillier, E.R.M., Collins, R.A. (2000) The contribution of replication orientation, gene direction, and signal sequences to base composition asymmetries in bacterial genomes. *Journal of Molecular Evolution*, **50**:249-257.

180. Valenzuela, C.Y. (1997) Non random DNA evolution. *Biology Research*, **30**:117-123.

181. Vologodskii, A.V., Frank-Kamenetskii, M.D. (1982) Theoretical study of cruciform states in superhelical DNAs. *FEBS Letters*, **143**:257-260.

182. Wada, K.-N., Doi, H., Tanaka, S.-I., Wada, Y., Furusawa, M. (1993) A neo-darwinian algorithm: asymmetrical mutations due to semiconservative DNA-type replication promote evolution. *Proceedings of the National Academy of Sciences of the United States of America*, **90**:11934-11938.

183. Watanabea, H., Gojobori, T., Miura, K.-I. (1997) Bacterial features in the genome of *Methanococcus jannaschii* in terms of gene composition and biased base composition in ORFs and their surrounding regions. *Gene*, **205**:7-18.

184. Watson, J.D., Crick, F.H.C. (1953) A structure for deoxyribose nucleic acid. *Nature*, **171**:737-738.

185. Wilquet, V., Van de Casteele, M. (1999) The role of the codon first letter in the relationship between genomic GC content and protein amino acid composition. *Research in Microbiology*, **150**:21-32.

186. Wright, S. (1969) Evolution and the genetics of populations. Volume 2, The theory of gene frequencies. The University of Chicago Press. ISBN 226-91050-4.

187. Wu, C.-I. (1991). DNA strand asymmetry. *Nature*, **352**:114-114.

188. Wu, C.-I., Maeda, N. (1987) Inequality in mutation rates of the two strands of DNA. *Nature*, **327**:169-170

189. Zharkikh, A; (1994) Estimation of evolutionary distances between nucleotide sequences. *Journal of Molecular Evolution*, **39**:315-329.

190. Zeigler, D.R., Dean, D.H. (1990) Orientation of genes in the Bacillus subtilis chromosome. *Genetics*, **125**:703-708.

191. Zuker, M. (1989) Computer prediction of RNA secondary structure. *Methods in Enzymology*, **180**:262-289.

CURRICULUM VITÆ

Etat civil

LOBRY, Raymond, Jean, 33 ans
Né le 01-07-66 à Grenoble (Isère)
Marié le 23-11-91 à Villeurbanne (Rhône)
Deux enfants

Diplômes

* Baccalauréat série D, mention TB, Nice, 1983
* Deug B1, mention TB, major, Strasbourg, 1984
* Deug B2, mention B, major, Strasbourg, 1985
* Licence de biochimie, mention AB, Lyon, 1986
* Prix scientifique Philips pour les jeunes, *Étude d'une classe particulière d'automate cellulaire*. Paris, 1986
* Maîtrise de Biotechnologie, mention B, major, Lyon, 1987
* DEA de Biométrie, mention B, major, Lyon, 1988
* Thèse soutenue le 9 septembre 1991 devant l'Université Claude Bernard, Lyon. Titre : *Ré-évaluation de modèle de croissance de Monod. Effet des antibiotiques sur l'énergie de maintenance*. Jury : A. Cheruy, J. Demongeot, J.-P. Flandrois, E. Jolivet, A. Pavé

Expériences professionnelles

* Cadre recherche, bioMérieux, 1991-1992
* Maître de Conférence Universitaire 2ème Classe depuis octobre 1992 à Lyon I.
* Maître de Conférence Universitaire 1ère Classe depuis septembre 1997 à Lyon I
* Consultant pour le groupe CREALIS (1997-1998)
* Membre de "l'Editorial Board" de *Applied & Environmental Microbiology* (1998-2001)

Liste des publications

Frank, A.C., Lobry, J.R. (2000) Oriloc: prediction of replication boundaries in unannotated bacterial chromosomes, *Bioinformatics*, in press.

Lobry, J.R. (1999) A nice wrong model for the evolution of DNA base frequencies. *Physica A*, **273**:100-103.

Lobry, J.R. (1999) Genomic landscapes. *Microbiology Today*, **26**:164-165.

Kano-Sueoka, T., Lobry, J.R., Sueoka, N. (1999) Intra-strand biases in bacteriophage T4 genome. *Gene*, **238**:59-64.

Frank, A.C., Lobry, J.R. (1999) Asymmetric substitution patterns : a review of possible underlying mutational or selective mechanisms. *Gene*, **238** :65-77.

Picardeau, M., Lobry, J.R., Hinnebusch, B.J. (1999) Physical mapping of an origin of bidirectional replication at the center of the *Borrelia burgdorferi* linear chromosome. *Mol. Microbiol*., **32**:437-445.

Lobry, J.R., Lobry, C. (1999) Evolution of DNA base composition under no-strand-bias conditions when the substitution rates are not constant. *Molecular Biology and Evolution*, **16**:719-723.

Charles, H., Mouchiroud, D., Lobry, J.R., Goncalves, I., Rahbe, Y. (1999) Gene size reduction in the bacterial aphid endosymbiont, Buchnera. *Molecular Biology and Evolution*, **16**:1820-1822.

Perrière, G., Lobry, J.R. (1998) Asymmetrical coding sequence repartition and codon adaptation index values between leading and lagging strands in seven bacterial species. *in* Proceedings of the first international conference on bioinformatics of genome regulation and structure (Novosibirsk, Russia, August 24-31, 1998) **2**:254-255.

Lobry, J.R. (1997) Influence of genomic G+C content on average amino-acid composition of proteins from 59 bacterial species. *Gene*, **205**:309-316.

Galtier, N., Lobry, J.R. (1997) Relationships between genomic G+C content, RNA secondary structures, and optimal growth temperature in prokaryotes. *Journal of Molecular Evolution*, **44**:632-636.

Perrière, G., Lobry, J.R., Thioulouse, J. (1996) Correspondence discriminant analysis: a multivariate method for comparing classes of protein and nucleic acids sequences. *Computer Applications in the Biosciences*, **12**:519-524.

Lobry, J.R. (1996) A simple vectorial representation of DNA sequences for the detection of replication origins in bacteria. *Biochimie*, **78**:323-326.

Lobry, J.R. (1996) Origin of replication of *Mycoplasma genitalium*. *Science*, **262**:745-746.

Lobry, J.R. (1996) Asymmetric substitution patterns in the two DNA strands of bacteria. *Molecular Biology and Evolution*, **13**:660-665.

Lobry, J.R. (1995) Properties of a general model of DNA evolution under no-strand bias conditions *Journal of Molecular Evolution*, **40**:326-330; **41**:680.

Thioulouse, J., Lobry, J.R. (1995) Co-inertia alaysis of amino-acid physico-chemical properties and protein composition with the ADE package. *Computer Applications in the Biosciences*, **11**:321-329.

Rosso, L., Lobry, J.R., Bajard, S., Flandrois, J.P. (1995) Convenient Model to Describe the Combined Effects of Temperature and pH on Microbial Growth. *Applied and Environmental Microbiology*, **61**:610-616.

Lobry, J.R. (1995) Unexpected behaviour of Monod's bacterial growth model. pp 149-154 *in* Mathematical population dynamic: analysis of heterogeneity (Arino, O., Axelrod, D., Kimmel, M. *eds*). Wuerz, Winnipeg.

Lobry, J.R., Gautier, C. (1994) Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 *Escherichia coli* chromosome-encoded genes. *Nucleic Acids Research*, **22**:3174-3180.

Rosso, L., Lobry, J.R., Flandrois, J.-P. (1993) An Unexpected Correlation between Cardinal Temperatures of Microbial Growth Highlighted by a New Model. *Journal of Theoretical Biology*, 162:447-463.

Lobry, J.R., Carret, G., Flandrois, J.-P. (1992) Maintenance requirements of Escherichia coli ATCC 25922 in the presence of sub-inhibitory concentrations of various antibiotics, *Journal of Antimicrobial Chemotherapy*, **29**:121-127.

Lobry, J.R., Flandrois, J.-P., Carret, G., Pavé, A. (1992) Monod's bacterial growth model revisited. *Bulletin of Mathematical Biology*, **54**:117-122.

Carret, G., Flandrois, J.-P., Lobry, J.R. (1991) Biphasic kinetics of bacterial killing by quinolones. *Journal of Antimicrobial Chemotherapy*, **27**:319-327.

Lobry, J.R., Flandrois, J.-P. (1991) Comparison of Estimates of Monod's growth model from the same data set. *Binary*, **3**:20-23.

Lobry, J.R., Rosso, L., Flandrois, J.P. (1991) A FORTRAN subroutine for the Determination of Parameter Confidence Limits in Non-linear Models. *Binary*, **3**:86-93.